Rui Carlos Pinto Borges

A Bayesian hierarchical robust model to detect and identify exonic CNVs

Department of Mathematics
Faculty of Sciences of Porto University
2016

Rui Carlos Pinto Borges

A Bayesian hierarchical robust model to detect and identify exonic CNVs

Thesis submitted to the Faculty of Sciences
of Porto University to obtain the Master
degree in Mathematical Engineering

supervised by:

Ana Paula de Frias Viegas Proença Rocha
Department of Mathematics
Faculty of Sciences of Porto University

and co-supervised by:

José Luís Silva Borges Costa
Institute of Molecular Pathology and Immunology of the Porto University
Medical Faculty of Porto University

Department of Mathematics
Faculty of Sciences of Porto University
2016

"*As you know from [...] introductory statistics, thirty is infinity.*"

Andrew Gelman

Acknowledgments

It is a pleasure to thank the many people who made this thesis possible.

I would like to gratefully acknowledge the guidance of my supervisor, Dr. Ana Paula Rocha, who has been abundantly helpful and has assisted me in numerous ways. She was the very first professor who encourage me in the very beginning of this journey, in which I barely could distinguish a vector from a scalar. I suspect that she was behind my increasingly interest in probabilities and simulation, the key ingredients of Bayesian Inference. The discussions I had with her were of endless value.

I would like to say a big thanks to Dr. José Luís Costa, who has come up with this problem and the data, and with, I would add, an enormous patience to discuss all the mathematical details.

I would like to say a big thanks to all my professors for the precious knowledge, the motivation, and for the corridor discussions, so many times, of extreme importance. Particularly, I would like to thank Dr. Rute Almeida for all the advices. I am also grateful to all my master colleagues for all the help, precious company and support.

My final words go equally to my family whom I want to thank for their guidance in whatever I pursue, and to my life partner Nádia Monteiro for her inestimable moral support, companionship and love.

Abstract

Copy number variation (CNV) of DNA fragments is a particular type of structural genetic variation. The characterization of CNVs has been an instrumental tool to understand different diseases. As an example, gene copy number alterations are frequent in cancer cells.

In this study, we present a Bayesian inferential framework to detect and identify CNVs the exonic regions of protein-coding-genes. Using the coverage readings from next-generation sequencing methodologies, we propose to standardize the case study coverage using a reference coverage. A hierarchical robust model with a normal likelihood is then implemented to assess the posterior distribution of each the exonic coverage ratio means. Markov Chain Monte Carlo via Hamiltonian Monte Carlo was performed to obtain random samples from the posterior distribution.

We present and discuss a case study with the *BRCA*1 and *BRCA*2 genes. The coverage standardization using a reference coverage is not affected by the existence of coverage readings from exonic CNVs. The use of a robust approach adds explanatory value to the model. The identification of CNVs proves straightforward using exon-specific Bayes factors. Values above 2 appear to be sufficient to properly identify true CNVs, but most important, eliminate artifact CNVs. In addition, an increased scale parameter in the distribution of the exonic coverage means, appears to be a promising pattern to detect CNVs, in a preliminary phase.

We conclude that the robust hierarchical Bayesian model satisfactory identifies and characterize exonic CNVs in protein-coding-genes. Our model was validated in the *BRCA*s genes, but may be generalized for any gene we might want to study.

Contents

List of Tables

List of Figures

Notation

Gelman et al. (2014a) notation is adopted in this thesis.

We opt for a compact notation format, that in some cases can be an abuse of the standard mathematical notation, it is more intuitive in others. $p(.)$ is interpreted as the probability of a specific event or the marginal probability distribution. $p(.|.)$ is the conditional probability density, with specified arguments on the right. Often conditioning quantities are implicit, however, conditioning on the data $p(.|y)$ will always be indicated.

Observations are symbolized as y, while the predictive data (future observations) are denoted as y^*.

Continuous and discrete distributions are equally treated.

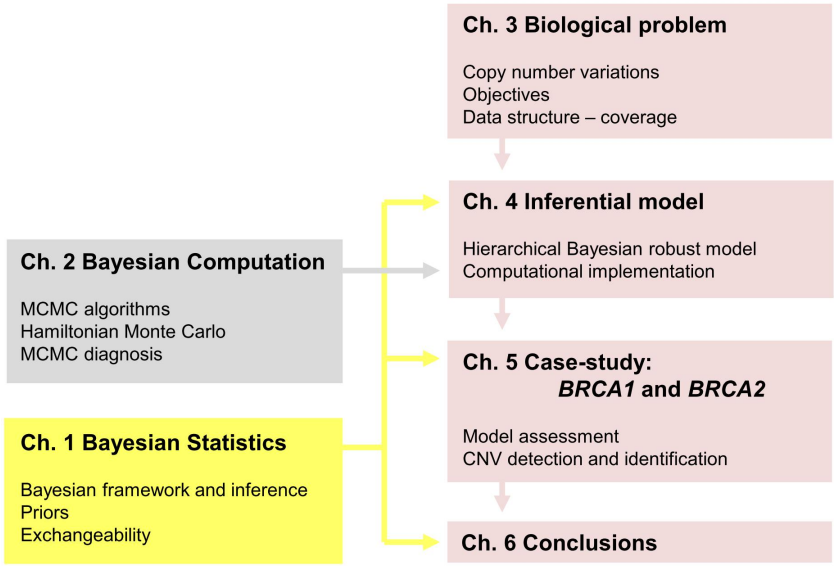
When standard probability distributions are used, a notation based on the name of the distribution is preferred. If θ has a normal density probability function with mean μ and variance σ^2, then we write $p(\theta|\mu, \sigma^2) = N(\theta|\theta, \sigma^2)$. Similarly, we use the notation $\theta \sim N(\mu, \sigma^2)$ to specify random variables.

Preface

The main objective of this thesis is to create an inferential model which permits to detect and identify CNVs on the exonic regions of genes. Such a problem requires both theoretical and applied considerations. We opt for a thesis organization that separates the theoretical background, necessary to understand the inferential approach and associated methodologies, from the biological problem, that includes the model construction and analysis.

Chapters 1 and 2 are meant to introduce the main concepts of the Bayesian inference and computation. Those who are familiar with the Bayesian inferential approach to probability and MCMC methodologies are free to skip these chapters.

Chapters 3, 4, 5 and 6 are devoted to the biological problem of CNV detection and identification. These core chapters include the data description, model building and assessment and model utility in the context of the biological problem.



Chapter 1

Introduction to Bayesian statistics

Bayes' solution to a problem of inverse probability was presented in *An Essay towards solving a Problem in the Doctrine of Chances* (Bayes and Price, 1763), published by the Royal Society in 1763 after Bayes' death. This essay contains a statement of a special case of the Bayes's theorem and can be considered the first historical milestone of Bayesian statistics (figure 1.1). Chapter 1 intends to give an overview of the fundamentals of Bayesian inference and statistics.

[376]

PROBLEM.

Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

Figure 1.1. Reverend Thomas Bayes (1702-1761) and an excerpt of the *An Essay towards solving a Problem in the Doctrine of Chances* (1763) where the problem of the inverse probability is stated.

1.1 Probability as uncertainty

In Bayesian statistics probability is understood as a measure of uncertainty (Bernardo, 2003). Bayesian approach uses prior knowledge of the study phenomenon together with the information that can be retrieved from the data, to estimate population parameters (Gelman et al., 2014a).

1.1.1 Bayesian knowledge

The introduction of prior information that may exist about the problem (including familiarity with the phenomenon or studies that have been carried out) is one of the major novelties of the Bayesian statistics (Gelman, 2002a). In fact, in most cases there are typically a set of data (even if sparse) or prior information about the process to be modeled. The general principle is simple, whatever the state of knowledge about some phenomenon, it can be expressed as a probability distribution (Bernardo, 2003).

The Bayesian statistics provides a natural method to introduce uncertainty accounting for experimental evidence: the knowledge concerned to the phenomenon or judgments formed prior to the random experience are mathematically expressed as a prior distribution, and information contained in data, belonging to a certain parametric family, are expressed as a likelihood function (Hoff, 2009). Combining the prior distribution and the likelihood function, we can obtain the posterior probability distribution, which expresses our uncertainty reviewed in the light of the data (Gelman et al., 2014a).

1.2 Bayesian Inference

In the Bayesian approach, the basis for the statistical procedure is the combination of any new probabilistic information to that one which is already available. Bayes' rule provides a formal framework to update beliefs in light of new information, being, as its name implies, the basis of the Bayesian inference (Gelman et al., 2014a).

Baye's theorem implicitly defines the conditional distribution of parameters given the data, i.e. the posterior distribution (Hoff, 2009). Consider that θ is an unobserved parameter (possibly a vector) and $y = (y_1, ..., y_n)$ an observed random sample of size n. The Bayes' rule states,

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta, y)}{\int_\Theta p(\theta, y) d\theta} = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta) d\theta}$$

The term in the denominator does not depend on θ, thus the posterior distribution can be

simply expressed by the product,

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Bayes' rule jointly integrates the prior information about the random experience and the sample information (Gelman et al., 2014a). Formally, the prior distribution $p(\theta)$ expresses the initial knowledge about θ prior observing y; the likelihood $p(y|\theta)$ describes the process giving rise to the data in terms of unknown θ; and the posterior distribution $p(\theta|y)$ is the probability recalculated based on the likelihood and the prior distribution, expressing what is known about θ after observing y (Hoff, 2009).

The posterior distribution gives us a refined estimate of θ (Gelman et al., 2014e). It can be showed that the prior variance can be decomposed in the terms,

$$V(\theta) = E[V(\theta|y)] + V[E(\theta|y)]$$

The expected posterior variance is on average lower than the prior variance by $V[E(\theta|y)]$, which corresponds to the information we expect the experiment to provide (Gelman et al., 2014e).

Bayesian inference becomes straightforward when the posterior is determined (directly or simulated). Inferences in Bayesian statistics are always based on a posterior probability distribution of θ, which includes point and interval estimates and also hypotheses testing.

1.2.1 Point estimates

In the Bayesian approach the posterior distribution is used to estimate the parameters under study. However, the estimators of θ must respect quality criteria. To obtain the quality status of the Bayesian estimators we use the loss function $L(\theta, a)$, which measures the loss of using the value a to estimate the parameter θ. The optimal estimate is chosen to minimize the expected loss $E[L(\theta, \hat{\theta})]$, calculated on the posterior distribution $p(\theta|x)$ (Ghosh et al., 2006).

Two types of loss functions are commonly used: the quadratic loss function ($L = (a - \theta)^2$) and the absolute loss function ($L = |a - \theta|$) (Carlin and Louis, 2009a). The posterior mean minimizes the posterior expected loss for the quadratic loss function. The posterior median minimizes the posterior risk for the absolute-value loss function. Other point etimators exist, as the posterior maximum (Carlin and Louis, 2009a). The posterior maximum finds the maximum of the posterior distribution an can be useful summarizing skewed distributions.

1.2.2 Interval estimates

In addition to point summaries, it is nearly important to report posterior uncertainty. A credible interval is an interval in the domain of the posterior probability distribution. A central interval for θ at $(1-\alpha)$ of posterior probability corresponds to the range of values (a, b) above and below which lies exactly $\alpha/2$ of the posterior probability (Ghosh et al., 2006).

$$p(a \leq \theta \leq b|y) \geq (1-\alpha)$$

Where α is a small positive number between 0 and 1.

A slightly different summary of the posterior parameter uncertainty is the highest posterior density region, defined as the region R that contains $(1-\alpha)$ of posterior probability, but in which the density within the region is never lower than that outside (Ghosh et al., 2006).

$$p(\theta_0|y) \geq p(\theta_1|y) \quad \forall \quad \theta_0 \in R, \quad \theta_1 \notin R$$

For unimodal, more-or-less symmetric distributions, highest posterior density based and quantile-based credible intervals would not be too different. But when the posterior densities become more complex (bimodal with well-separated modes) the highest posterior density credible region are two disjoint intervals, whereas the central quantile-based credible region is a single interval by construction (Gelman et al., 2014e).

In Bayesian inference it makes sense to enunciate the Bayesian credibility intervals (or regions), as a state of probability. A legitimate interpretation of Bayesian credibility regions is that, given the information we observed, we are $(1-\alpha)$ confident that the true value of θ is within the obtained interval (or region) (Ghosh et al., 2006).

1.2.3 Hypothesis testing

The Bayesian approach to hypothesis testing is quite simple and intuitive. Consider the hypothesis testing problem:

$$H_0 : \theta \in \Theta_0 \quad \textit{vs.} \quad H_1 : \theta \in \Theta_1$$

Where Θ_0 and Θ_1 are compound hypothesis. We can define the posterior probability of each hypothesis as λ_0 and λ_1:

$$\lambda_0 = p(\theta \in \Theta_0|y) \quad \text{and} \quad \lambda_1 = p(\theta \in \Theta_1|y)$$

λ_0 and λ_1 can be used to conclude which hypothesis must be considered. If $\lambda_0 > \lambda_1$ we have favorable posterior information to accept H_0, on the contrary, if $\lambda_1 > \lambda_0$ the hypothesis H_1 is more likely and we should reject H_0. To accept or to reject one of the hypotheses the ratio

between λ_0 and λ_1 probabilities is commonly used: the Bayes factor (BF). (Ghosh et al., 2006).

$$BF = \frac{\lambda_0}{\lambda_1} = \frac{p(\theta_0|y)}{p(\theta_1|y)} = \frac{p(y|\theta_0)}{p(y|\theta_1)} \frac{p(\theta_0)}{p(\theta_1)}$$

The Bayes factor is used considering the Jeffrey criterion, which state that the hypothesis with an odd ratio higher than 1 should be favored (Ghosh et al., 2006). The Bayes factor can be also used as a degree of confidence: a Bayes factor bigger that 10 is generally seen as strong evidence while bigger that 20 is decisive evidence for the numerator model/hypotheses.

The Bayesian hypothesis testing can include a loss function $L(\theta)$, which integrates the posterior risk. Using a risk function in the Bayes factor to determine the best choice among several is nothing else than a probabilistic optimization problem. These type of problems arise naturally in the Bayesian framework, when more than probabilities are needed to be accounted (money, time, security, efficiency, ...).

1.2.4 Predictive distribution

Predictive distributions are widely used in Bayesian statistics during the model assessment phase. We have been using the posterior distribution $p(\theta|y)$ broadly in inference, but we can be interested in make predictive probabilistic statements about an unobserved quantity (a future observation) y^*, considering the obtained observed data $y = (y_1, ..., y_n)$. Predictive inferences rely in the posterior predictive distribution $p(y^*|y)$, which is a conditional distribution on the observed values (Gelman et al., 2014g).

The posterior predictive distribution can be defined using the posterior distribution,

$$p(y^*|y) = \int p(y^*|\theta)p(\theta|y)d\theta$$

which implies that future data is independent of past data, conditional on the parameters (Gelman et al., 2014g). The posterior predictive distribution is obtained by integrating the product of the data model distribution $p(\theta|y)$ with the posterior distribution with respect to the model parameters $p(y^*|\theta)$. In consequence, the posterior predictive distribution has the same mean as the posterior distribution, but a greater variance, due to the additional sampling uncertainty (Gelman et al., 2014g; Gelman and Shalizi, 2013).

In some cases the form of $p(y^*|y)$ can be derived directly, but it is often easier to sample from $p(y^*|y)$ using Monte Carlo iterates (say L) that we might have from the posterior distribution: a random vector of parameters θ_l can be obtained from the posterior distribution and used to simulate y^* according to the $p(y^*|\theta_l)$ distribution. $y^* = (y_1^*, ..., y_l^*, ..., y_L^*)$ is a random sample from $p(y^*|y)$ (Albert, 2009a).

1.3 Prior distribution

Bayesian inference includes two sources of assumptions, the likelihood and the prior distribution. The use of a prior distribution caused distrust and criticism from other probability schools (Kass and Wasserman, 1996), being known as subjective for several decades (Gelman et al., 2014a). However, this sobriquet was unfairly received since any type of probability statement is subjective. There are many aspects to take into account choosing a prior distribution: some are choose for mathematical convenience, others to express lack of information, and others to specify information about parameters (Kass and Wasserman, 1996). Thus, the prior distribution may be as generic and uninformative as one wants, or alternatively, integrate data from past experiences (Gelman, 2002b), making the whole Bayesian process objective.

In many problems there is no relevant information on the parameters of interest and a non-informative prior should be considered. When some information is available a informative prior could be used, but the prior should not be too stringent, because it can limit the posterior distribution to evaluate the prior zero density regions. A good property of priors is that the influence of the prior generally goes to zero as we collect more data, given prevalence to the observational information (Kass and Wasserman, 1996).

1.3.1 Conjugate priors

A prior is conjugate for a family of distributions if the prior and the posterior are of the same distributional family (Gelman et al., 2014e). Exponential families have conjugate priors in general (Kass and Wasserman, 1996), in which the likelihood, considered in terms of the parameters, has a kernel in the same form as the prior distribution.

Conjugate prior are preferable because they generally simplify obtaining posterior distributions and/or obtaining marginal/conditional posterior distributions (Kass and Wasserman, 1996). In addition, the conjugate priors can be either informative or non-informative, depending on the chose parameter values (Gelman et al., 2014e).

1.3.2 Non-informative priors

A prior distribution is non-informative if the prior is flat relative to the likelihood function, returning that, it has minimal impact on the posterior distribution of θ (Gelman, 2002b). The uniform distribution or the normal distribution with variance large enough, are good examples of non-informative priors.

An additional care with non-informative prior must be considered: the propriety property

(Kass and Wasserman, 1996). A prior $p(\theta)$ is improper if,

$$\int_{\Theta} p(\theta)d\theta = \infty$$

The propriety property implies that the kernel of the prior must integrate to a finite number. Improper priors may result in proper or improper posteriors, but in the last case, we are not in conditions to proceed to inference (Kass and Wasserman, 1996). The use of improper priors require the additional effort of knowing if the posterior is proper.

Improper priors are often used because they are generally non-informative. In practical cases, the improper priors use can be justified when the data are informative enough about the parameter of interest, thus we do not need specifying our ignorance exactly (Gelman et al., 2014e). If there are lack of data, the use of a improper prior is not recommended (Kass and Wasserman, 1996).

1.3.3 Informative priors

Informative priors are not dominated by the likelihood, and therefore have a considerable impact on the posterior (Gelman, 2002b).

Usually the main challenge of using informative priors is to choose a reasonable distributional family. Some choices are obvious: a normal or a t-student distribution are well suited for real parameters; gamma, inverse-gamma and log-normal are adequate for precision parameters and variance components; beta distribution fits parameters ranging in the $[0, 1]$ interval (Gelman et al., 2014e). Once a prior family, and particular parameter values are chosen (mean, variance, shape, scale), it is recommended to verify if the statistical summaries of the prior are consistent with our prior beliefs (Gelman et al., 2014e).

1.4 Interpretations of probability

The Bayesian and Classical methods (meaning Frequentist) collide mainly in the way they interpret the parameters and data that are subjected to inference (Lindley, 2000). Classic inference is represented by Neymann, Pearson and Wald and respects the orthodox view that sampling is infinite and decision rules can be sharp. Data is seen as a repeatable random sample (there is a frequency), in which the underlying population parameters remain constant during the repeatable process (Gelman and Shalizi, 2013). Bayesian appears with Bayes, Laplace and de Finneti, and is characterized by threatening probabilistically any unknown quantity, which can always be updated. Data are finite and observed from a realized sample, and the parameters are considered unknown quantities, described as probability distributions (Bernardo, 2003).

Bayesian approach considers the parameters as random variables while the Classical approach considers the parameter as fixed, even though unknown. On the contrary, while data is fixed in Bayesian, it is a random sample for Frequentist (Lindley, 2000). Table 1.1 summarizes some of the main aspects that differentiate Bayesian and Classical statistics.

Topics	**Classic**	**Bayesian**
Probability	Limit of empirical frequencies	Subjective belief
Parameter	θ is fixed	θ is random
Estimation	Likelihood based	Posterior based
Sources of information	Data only	Data and prior beliefs
Inference	Interpreted in terms of the long-run behavior of y	Interpreted as probability statements about θ
Computation	Optimization	Integration
Uncertainty	Often based on asymptotics	Exact

Table 1.1. Comparative aspects of the Classic and the Bayesian types of inference: schematic summarization of the topics discussed in Lindley (2000).

1.4.1 Advantages and disadvantages of Bayesian Statistics

The Bayesian statistics has both advantages and disadvantages in its implementations and uses.

One of the main advantage of the Bayesian inference is that it is simple in principle and provides a framework for coherent inference based on the posterior distribution, which integrates prior beliefs (various sources of information, including constraints) and additional information (Bernardo, 2003). Bayesian approach obeys to the likelihood principle being conditional on the observed data (Berger and Wolpert, 1988), and not in the data that were possible but not observed. As consequence, inference for small samples is always exact. Bayesian results often have good Classical properties, being Classical analysis a special case of the Bayesian under a particular prior in some cases (Lindley, 2000).

Because probability is seen as a mean to describe uncertainty on the parameters, Bayesian inference naturally deals with conditioning, marginalization and decision theory (Bernardo, 2003). Bayesian inference naturally penalizes complex models, but complex models can be easily constructed using hierarchical modeling (Gelman et al., 2014g). Hierarchical Bayes deals with multiple testing inherently.

Modern computational techniques facilitate to work with Bayesian models (Gelman et al.,

2014b). Partly because of the recent developments in simulation science, it is possible to actually sample (directly or indirectly) from the posterior, to perform inference.

One of the main disadvantages of the Bayesian inference regards the posterior distribution. Although Bayesian inference provides a simple and well established framework to compute the posterior, it is often difficult and time-consuming to implement it in practice (Gelman et al., 2014b). For complicated posteriors, model exploration, assessment and comparison can be compromised due to analytical and numerical constrains. In this respect, Classical inference is simpler in standard statistical analysis (Lindley, 2000). Moreover, Bayesian inference is conditional on the observed data and may not be generalize.

1.5 Exchangeability

A sequence $(y_1, ..., y_n)$ of random variables is finitely exchangeable if the joint distribution $p(y_1, ..., y_n)$ is invariant under any permutation of the indexes of the random variables,

$$p(y_1, ..., y_n) = p(y_{(1)}, ..., y_{(n)})$$

for all permutations on the set $\{1, ..., n\}$ (Bernardo, 1996). An infinite sequence is infinitely exchangeable if any finite sub sequence is finitely exchangeable (Bernardo, 1996).

The assumption of exchangeability does not mean that the observations are similar, but instead that there are no information to specifically distinguish any observation of another. Exchangeable observations cannot be grouped or ordered by principle (Gelman et al., 2014f). In Bayesian statistics the exchangeability is a typical probabilistic assumption (Bernardo, 1996; Good, 2002).

By definition, independent and identically distributed random variables are exchangeable, but the contrary is not verified (Good, 2002). Consider the Polya's urn, containing two black balls and three white. We know that the probability of any of the ball configurations is not independent of the last retrieval, but they are exchangeable, because each ball color configuration has the same probability.

$$p(1, 1, 0, 0, 0) = \frac{1}{10} = p(0, 1, 0, 0, 1)$$

The standard bi-dimensional normal distribution is a good example of exchangeable variables: $p(x, y) = \mathrm{N}(0, 0, 1, 1, \rho_{xy})$. x and y can be exchanged, obtaining the the same probabilistic inferences, however both are by definition not independent.

Chapter 2

Bayesian computation

The Bayesian computation provides a framework to obtain meaningful summaries of the posterior distribution. Chapter 2 presents some of the most currently used simulation methods in Bayesian Computation (figure 2.1), particularly focusing the iterative Monte Carlo simulation techniques.

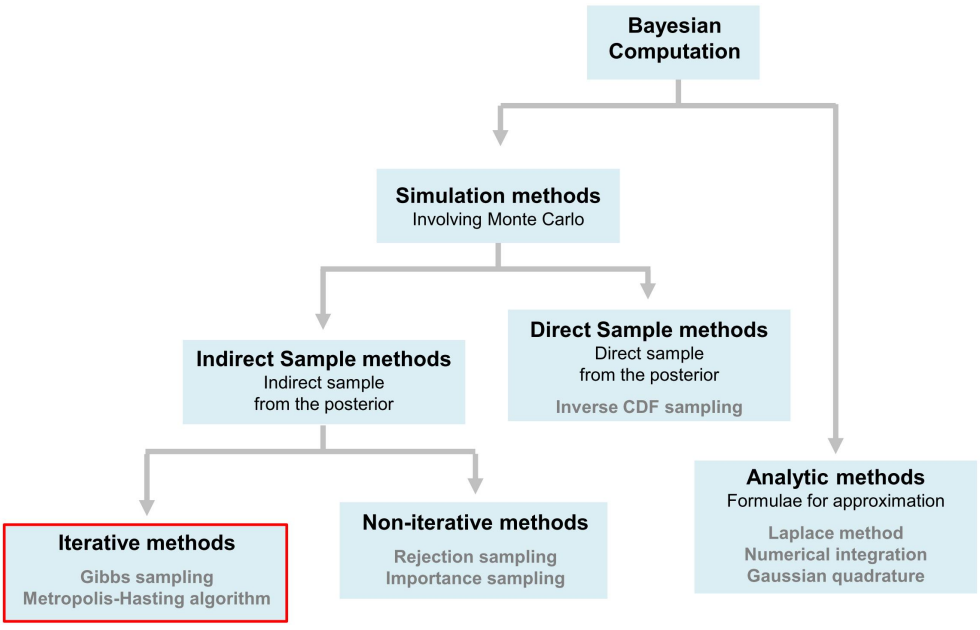


Figure 2.1. Schematic view of the Bayesian computation techniques.

2.1 Monte Carlo simulation

Approximations to integrals can be carried out by Monte Carlo integration (Gamerman and Lopes, 2006a). Suppose θ has a density $p(\theta)$ and its intended to calculate the quantity $\gamma = E[h(\theta)]$.

$$\gamma = E[h(\theta)] = \int h(\theta)p(\theta)d\theta$$

Consider $\{\theta^1, ..., \theta^L\}$, L independent and identically distributed samples from the distribution $p(\theta)$, then the estimator,

$$\hat{\gamma} = \frac{1}{L}\sum_{i=1}^{L} h(\theta^i)$$

converges to $E[h(\theta)]$ almost surely as $L \to \infty$, by the strong law of large numbers. Naturally, as L increased, the quality of the approximation increases (Gamerman and Lopes, 2006a).

$$V[\hat{\gamma}] = \frac{1}{L}V[h(\theta)]$$

In Bayesian inference, $p(\theta)$ is the posterior distribution $p(\theta|y)$ and thus $E[h(\theta)]$ is the posterior mean of $h(\theta)$. In order to use the Monte Carlo approximation in Bayesian computation is only required to sample from the posterior distribution a sample of size L (Gamerman and Lopes, 2006a). Another facility of the Monte Carlo method is that any other quantity than the mean, can be similarly estimated, which includes probabilities (integration of the posterior density over an interval) using empirical proportions, quantiles using empirical quantiles or second moments (dispersion measures) (Albert, 2009b).

2.2 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods set up a Markov chain whose stationary distribution is the posterior distribution. These methods are called Monte Carlo because rely on random draws from the approximate distribution, and Markov chain because samples are drawn iteratively as a Markov chain (Albert, 2009c).

2.2.1 Gibbs sampling

Suppose we have a collection of k random variables denoted by $\theta = (\theta_1, ..., \theta_k)$. We assume that the full conditional distributions

$$\{p(\theta_i|\theta_j, j \neq i), i = 1, ..., k\}$$

are available for sampling and that we have some method to generate samples from the conditional distributions (Gamerman and Lopes, 2006b). It is not required the one dimensional conditional distributions $p(\theta_i|\theta_j, j \neq i)$ to have a close form, but for sampling it is

necessary to be able to write them up to a normalization constant (Gelman et al., 2014b). Furthermore, the conditional distribution of one variable given all others is proportional to the joint distribution (Gamerman and Lopes, 2006b),

$$p(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k) = \frac{p(\theta_1, ..., \theta_k)}{p(\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k)} \propto p(\theta_1, ..., \theta_k)$$

The idea behind the Gibbs sampling is very simple and intuitive. The algorithm proceeds as follows (Gamerman and Lopes, 2006a; MacKay, 2005b):

1. Generate a set of arbitrary starting values $(\theta_1^0, ..., \theta_k^0)$.
2. Perform random draws from the uni-variate conditional distributions.
   - Draw θ_1^1 from $p(\theta_1|\theta_2^0, ..., \theta_k^0, y)$
   - ...
   - Draw θ_k^1 from $p(\theta_k|\theta_1^1, ..., \theta_{k-1}^1, y)$
3. Repeat the process L times.
   The last sample would be $(\theta_1^L, ..., \theta_k^L)$.

We are interested in samples from the joint posterior distribution $p(\theta|y)$ but Gibbs provide draws from each of the univariate conditional distributions $p(\theta_i|\theta_j, y, i \neq j)$. It can be proved that $(\theta_1^l, ..., \theta_k^l)$ converges in distribution to $p(\theta_1, ..., \theta_k|y)$ as $L \to \infty$ being the convergence geometric in t (Gamerman and Lopes, 2006b).

Obtaining the marginal densities can be straightforward, specially if they all belong to the exponential family. However, in other cases we cannot be able to identified the conditional of some parameters as a standard distribution, making the sample process impossible. In these cases any non-iterative sampling algorithms can be used (eg. rejection sampling). Any valid way of generating samples from $\theta_i|y, \theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k$ will give a legitimate MCMC algorithm (Gamerman and Lopes, 2006b).

2.2.2 Metropolis-Hastings algorithm

Similar to the Gibbs sampling, the Metropolis-Hastings algorithm is an MCMC method. Metropolis-Hastings algorithm was firstly developed by Metropolis in 1953, being adapted for statistical analysis by Hastings in 1970 (Gamerman and Lopes, 2006c).

Suppose we pretend to sample from the posterior distribution $p(\theta_1, ..., \theta_k|y) = p(\theta|y)$. Let θ^t be the current parameter vector and $p(\theta|\theta^t)$ be a proposal distribution such that $p(\theta_a|\theta_b) = p(\theta_b|\theta_a)$. The Metropolis algorithm generate random samples as follows (Carlin and Louis, 2009b; Gamerman and Lopes, 2006c):

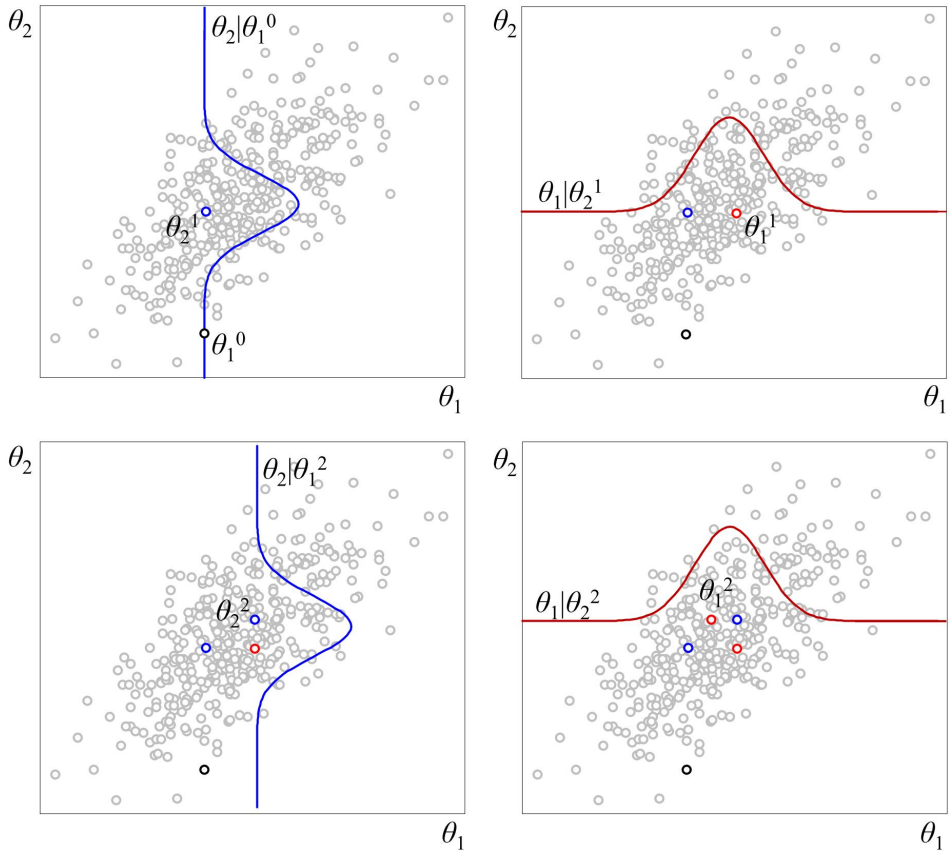


Figure 2.2. Gibbs sampling involves estimating a joint probability distribution of two or more random variables (here with θ_1 and θ_2), by sampling from conditional distributions. Based on MacKay (2005b).

1. Draw θ^* from $p(\theta|\theta^t)$, where θ^t is the current state of the chain.
2. Calculate the acceptance probability

$$\alpha(\theta^t, \theta^*) = \frac{p(\theta^*|y)}{p(\theta^t|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^t)p(\theta^t)}$$

3. Set $\theta^{t+1} = \theta^*$ with probability α, otherwise set $\theta^{t+1} = \theta^t$.

Note that even when a proposal is not accepted, it counts as an iteration and the last value is repeated as part of the sample (Gelman et al., 2014b). Metropolis algorithm always accepts when the posterior density of the proposal is higher than the density of the current value, but also accepts with probability less than 1 when the proposal density is lower (Gamerman and Lopes, 2006c). Normal distributions or t-student distributions are common choices for the proposal distribution (Gelman et al., 2014b). One can see that Metropolis algorithm is a stochastic version of an optimization algorithm, however, the higher or lower ability to move to places with lower density in the parameter space is critical in exploring the posterior distribution (which can contain local maxima) (Gamerman and Lopes, 2006c).

In certain regularity conditions it can be proved that $\theta^{(t)}$ converges in distribution to $p(\theta|y)$ (Gamerman and Lopes, 2006c). However, it must be stressed that the sequence $\theta^{(t)}$ is not independent.

A simple, but important generalization of the Metropolis algorithm was provided by Hastings in 1970 (Carlin and Louis, 2009b). Hastings drops the requirement of symmetry of the proposal distribution $p(\theta|\theta^t)$, and redefines the acceptance ratio as

$$\alpha(\theta^t, \theta^*) = min\left\{1, \frac{p(\theta^*|y)}{p(\theta^t|y)}\frac{p(\theta^t|\theta^*)}{p(\theta^*|\theta^t)}\right\}$$

With this modification, it can be shown that this algorithm converges to the required posterior distribution for any candidate proposal density $p(\theta|\theta^t)$ (Gamerman and Lopes, 2006c).

The proposal distribution can be anyone, but often normal or t-student proposals centered at the current iterate θ^t, are used (Gelman et al., 2014b). Asymmetric proposal such as gamma distributions, may also be used but require a Hastings adjustment. One can use proposal distributions that does not depend on the current parameter value. These distributions are known as independent samplers and require a Hastings adjustment (Albert, 2009c).

The Gibbs sampler can be seen as a special case of the Metropolis-Hastings. If the conditional distribution is used as the proposal distribution, the draws will be always accept (Gelman et al., 2014b). In the Gibbs sampler the proposal distribution $p(\theta_j^*|\theta_j^t, y)$ is such that $\theta_{-j}^* = \theta_{-j}^t$, which means that the others components of θ except θ_j do not change. Given that the acceptance probability,

$$\begin{aligned}
\alpha(\theta^*, \theta^t) &= \frac{p(\theta^*|y)p(\theta^t|\theta^*)}{p(\theta^t|y)p(\theta^*|\theta^t)} \\
&= \frac{p(\theta^*|y)p(\theta_j^t|\theta_{-j}^t, y)}{p(\theta^t|y)p(\theta_j^*|\theta_{-j}^t, y)} \\
&= \frac{p(\theta_j^*|\theta_{-j}^t, y)p(\theta_{-j}^t|y)p(\theta_j^t|\theta_{-j}^t, y)}{p(\theta_j^t|\theta_{-j}^t, y)p(\theta_{-j}^t|y)p(\theta_j^*|\theta_{-j}^t, y)} \\
&= 1
\end{aligned}$$

A very common approach is to combine the Metropolis-Hastings algorithm with the Gibbs sampler in building blocks (Gelman et al., 2014b; Gamerman and Lopes, 2006c). The Metropolis-Hastings steps are used when a particular posterior conditional distribution does not have a simple form to sample, and Gibbs sampler for the others. The Metropolis-Hastings can be single-component or blocked depending on the number of parameters to be sampled. For models with strongly dependent parameter the Metropolis-Hasting algorithm can be preferred to improve mixing, even though the Gibbs sampler could be implemented

(Gamerman and Lopes, 2006c). The tradeoff is between improving independence between iterates in Metropolis-Hastings and sampling from the proper conditionals with 100% of acceptance probability in Gibbs (Gelman et al., 2014b).

2.3 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo borrows an idea from physics to suppress the local random walk behavior commonly observed in the Gibbs and Metropolis-Hastings algorithms, thus allowing the Markov chain to move much more rapidly through the target distribution (Neal, 2011a). The random walk behavior is particularly limitative in complicated high-dimensional target distributions (Gelman et al., 2014c).

Let λ be the model parameters vector. For each component λ_j in the target space, the Hamiltonian Monte Carlo adds a momentum variable ϕ_j. Both λ and ϕ are updated together in a new Metropolis algorithm, in which the jumping distribution for λ is largely determined by ϕ (Neal, 2011a). Each iteration proceeds via several steps, during which the position and momentum evolve based on rules imitating the Hamiltonian dynamics (Gelman et al., 2014c). Hamiltonian Monte Carlos is also called hybrid Monte Carlo because it combines MCMC and the determinism of the Hamiltonian trajectory.

2.3.1 Momentum distribution

In Hamiltonian Monte Carlo, the posterior distribution is augmented by an independent distribution $p(\phi)$ on the momenta (Gelman et al., 2014b), thus defining a joint distribution,

$$p(\theta, \phi|y) = p(\phi)p(\theta|y)$$

We simulate from the joint distribution but we are only interested in the simulation of θ (the position vector). The vector ϕ has the same dimension as θ and is an auxiliary variable, introduced to allow the algorithm to move faster through the parameter space (Neal, 2011a).

It is common to give ϕ a multivariate normal distribution with mean 0 and covariance setted to a predefined mass matrix M (Gelman et al., 2014b). The mass matrix is so called by analogy to the physical model of Hamiltonian dynamics (Neal, 2011a). It commonly used a diagonal mass matrix, with $\phi_j \sim \mathrm{N}(0, M_{jj})$ (Gelman et al., 2014b).

2.3.2 Hamiltonian dynamics

The main part of the Hamiltonian Monte Carlo is a simultaneous update of the (θ, ϕ), performed by a discrete mimicking of the physical dynamics (Neal, 2011a). One common

discretized scheme for the Hamiltonian dynamics is the leapfrog method (Gelman et al., 2014b). Consider M the mass matrix, the covariance of the momentum distribution $p(\phi)$ and ϵ, a scaling factor that should be settled considering the number of leapfrog iterates L on the Hamiltonian dynamics step: $\epsilon L = 1$ (Neal, 2011a). The leapfrog discretization of the Hamiltonian dynamics becomes,

$$\begin{cases} \phi \leftarrow \phi + \frac{\epsilon}{2} \frac{d \log p(\theta|y)}{d\theta} \\ \theta \leftarrow \theta + \epsilon M^{-1} \phi \end{cases}$$

Lets explore some intuitive ideas about these equations. Suppose the algorithm move toward an area of low posterior probability. $\frac{d \log p(\theta|y)}{d\theta}$ will be negative in this direction, inducing a decrease in the momentum in the direction of movement. As the leapfrog steps continue to move into an area of lower density in θ-space, the momentum continues to decrease. A good property of the leap frog step is that if the iterations continue to move in the direction of decreasing density, the leapfrog step will slow to zero and then back down or curve around the dip (Gelman et al., 2014b).

2.3.3 HMC algorithm

Hamiltonian Monte Carlo algorithm uses the last iterated parameter vector θ^t and a mass matrix M as input. The Hamiltonian Monte Carlo algorithm comprises 3 main steps (MacKay, 2005a):

1. The momentum vector ϕ^t is updated. Usually, $\phi \sim \mathrm{N}(0, M)$
2. The discrete Hamiltonian dynamics is them performed, involving L leapfrog steps, each scaled by a factor ϵ.
   (a) Initial half-step of ϕ
   $$\phi \leftarrow \phi + \frac{\epsilon}{2} \frac{d \log p(\theta|y)}{d\theta}$$
   (b) $L - 1$ alternated updates of θ and ϕ
   $$\phi \leftarrow \phi + \epsilon \frac{d \log p(\theta|y)}{d\theta} \quad \wedge \quad \theta \leftarrow \theta + \epsilon M^{-1} \phi$$
   (c) Final half-update of ϕ.
3. The updated parameter and momentum vectors (θ^*, ϕ^*) are compared with the initial iterates (θ^t, ϕ^t) by an accept-reject step, calculating the acceptance probability.
   $$\alpha = \frac{p(\theta^*|y)p(\phi^*)}{p(\theta^t|y)p(\phi^t)}$$
   Set $\theta^{t+1} = \theta^*$ with probability $\min(\alpha, 1)$ and reject otherwise, $\theta^{t+1} = \theta^t$.

2.4 Convergence and Mixing

The major pitfalls in using MCMC methods are convergence and mixing. Convergence is achieved when the MCMC iterates can be safely thought of as coming from the stationary posterior distribution (Carlin and Louis, 2009b). Mixing is achieved when independent MCMC chains cover the same distribution (Carlin and Louis, 2009b).

Achieving convergence and mixing can be difficult: the iterations have not proceeded long enough and the simulations may be grossly unrepresentative of the target distribution; the within-sequence correlation can be high making simulation inferences less precise; and the presence of local maximums, can lead to cover different distributions (Gelman et al., 2014b).

These special problems of iterative simulation are handled by three good practices in inferential simulation:

- Discarding early iteration of the simulation runs
  To diminished the influence of the starting values, generally the first half of each sequence is discarded (Gelman et al., 2014b). Inferences will be based on the assumption that the distribution of the simulated values ψ_t for large enough t, are close to the target distribution $p(\theta|y)$ (Carlin and Louis, 2009b). The practice of discarding early iterations in Markov chain simulations is referred as *burn-in* (Neal, 2011b).

- Dependence of the iteration in each sequences
  MCMC iterates are not independent and successive samples are correlated (Gelman et al., 2014b). Autocorrelation will be reduced by thinning the sequences by keeping every k-th simulation draw from each sequence $\{\psi_i, \psi_{i+k}, \psi_{i+2k}, ...\}$ (Carlin and Louis, 2009b). In problems with large numbers of parameters, where computer storage is a problem, it can be useful to skip iterations.

- Multiple sequences with over-dispersed starting values
  Most of the recommended approaches to assess convergence of iterative simulation is based on comparing different simulated sequences, because to evaluate convergence and mixing, more than one independent sequence are needed (Carlin and Louis, 2009b). Thus, it becomes a good practice to simulate independent sequences (at least two), with starting points drawn from an overdispersed distribution (Neal, 2011b).

Visual representation of MCMC chains can be useful informing if the iterates have converged to its stationary distribution or if they are mixed, i.e. covering the same distribution (figure 2.3). The recommended practice is to combine formal and visual methods to assess the convergence of the MCMC chains (Gelman et al., 2014b; Neal, 2011b).

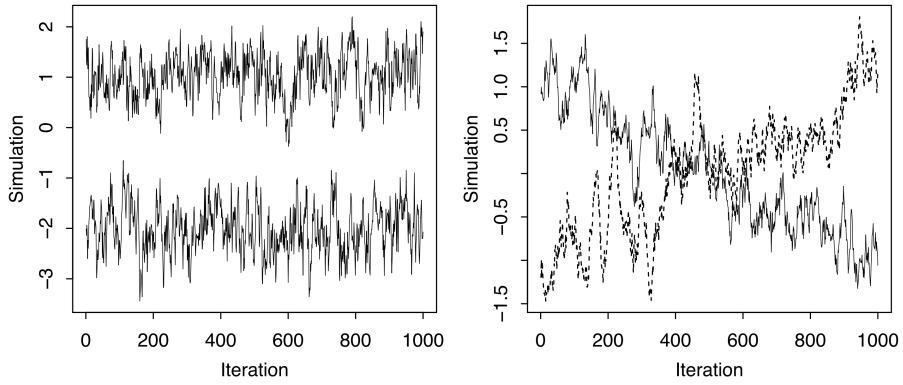


Figure 2.3. Convergence and mixing of two MCMC chains. In the left plot, while both sequences look stable, their non-overlapping suggest they have not converged to a common distribution. In the right plot, the two sequences cover a common distribution but none of sequences are stationary. Based on (Gelman et al., 2014b)

2.4.1 Scale reduction factor

One of the most common methods for monitoring MCMC convergence is the potential scale reduction factor ($\hat{R}$) proposed by Gelman and Rubin (1992). Multiple MCMC sequences are started from overdispersed initial points and compared. $\hat{R}$ assess the convergence of the chains comparing the variance and mean of each chain to the variance and mean of the combined chain (Gelman and Rubin, 1992).

Consider m parallel chains with $2n$ samples each. Only the last n better converged samples from each chain are used. The between-chain variance and pooled within-chain variance are defined as, respectively

$$\frac{B}{n} = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{j\cdot} - \bar{\psi}_{\cdot\cdot})^2 \qquad W = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{i=1}^{n}(\psi_{ji} - \bar{\psi}_{j\cdot})^2$$

where $\bar{\psi}_{j\cdot}$ is the mean of the samples in chain j and $\bar{\psi}_{\cdot\cdot}$ is the mean of the combined chains. Finally an estimate of $\hat{R}$ is obtained by dividing the pooled posterior variance with the pooled within chain variance,

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}} \quad \text{where} \quad \hat{V} = \frac{n-1}{n}W + \left(1 + \frac{1}{m}\right)\frac{B}{n}$$

If the chains have converged, $\hat{R}$ is close to 1, which makes it a useful indicator of convergence (Gelman and Rubin, 1992). $\hat{R}$ has some limitations (Venna et al., 2003): the chains might not have covered the state space and might discover new areas of high probability; it does not take higher-order moments into account (only the mean and variance); and it is applicable to only one variable at a time.

Chapter 3

CNV detection and identification

Copy number variations (CNVs) are genetic structural alterations of different types and sizes. In this thesis we focused the problem of detecting and identifying exonic CNVs. Exons corresponds to the portions of a gene that code for a protein (figure 3.1). Chapter 3 introduces some biological aspects of the CNV detection and identification problem and additionally discusses some aspects of the data structure.

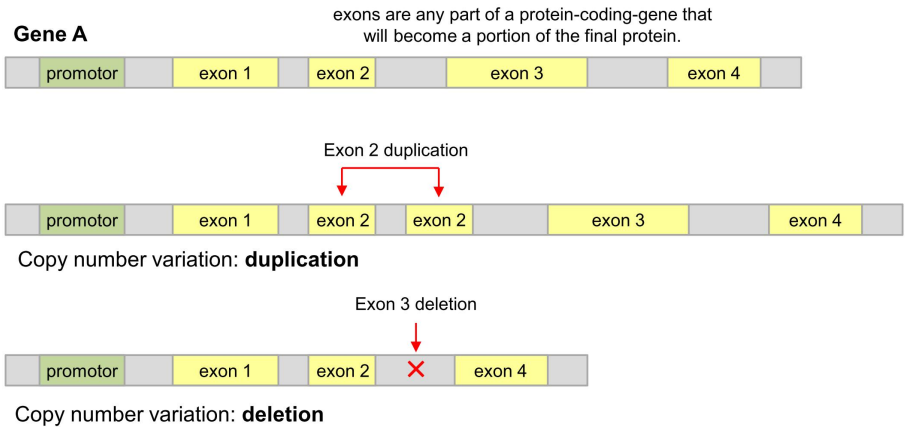


Figure 3.1. Schematic view of exonic structural alterations. A normal individual is compared with two copy number variants (duplication and deletion) for the exonic regions of the gene A.

3.1 Copy number variations

During the last years geneticists have been characterizing hundreds of repetitive regions in DNA (Redon et al., 2006), contributing to the idea that structural-level alterations are as important as the sequence-level genomic diversity, to understand the variability among individuals and populations (Stankiewicz and Lupski, 2010). These repetitive regions are denominated CNVs (copy number variations) and represents an imbalance between two genomes from one species, affecting segments of DNA ranging from one kilobase to several megabases in size (Feuk et al., 2006). CNVs account for approximately 12% of the human genomic DNA and are caused by structural rearrangements: deletions, duplications, triplications, insertions, or translocations can result in CNVs (Redon et al., 2006; Stankiewicz and Lupski, 2010).

Gene-specific CNVs have been associated with susceptibility and resistance to diseases. *CCL3L1* is a potent human immunodeficiency virus (HIV) suppressive chemokine and was shown that lower *CCL3L1* copy numbers were associated with markedly enhanced HIV/acquired immunodeficiency syndrome susceptibility (Gonzalez et al., 2005). The development of systemic autoimmunity is also associated with low *FCGR3B* copy numbers (Fanciulli et al., 2007). Confirmed *de novo* CNVs were significantly associated with more complex diseases, as autism (Sebat et al., 2007).

Next-generation sequencing technique is facilitating an increase in the efficiency and resolution of CNVs detection, becoming the most popular strategy (Metzker, 2010). CNVs are identified in two steps (figure 3.2): a sequencing step, in which the sample DNA is shredded into small fragments that are partially sequenced (reads); and an assembly step, in which the reads are aligned considering a reference sequence. In comparison with alternative methods, next-generating sequencing approach has the advantage of having higher resolution, more accurate estimation of copy numbers and higher capacity to identify novel CNVs (Alkan et al., 2011; Meyerson et al., 2010).

3.1.1 Coverage readings

A raw dataset from next-generation sequencing procedures includes the coverage readings and the respective genomic position. The coverage readings constitute the number of times a specific genomic site in the reference genome is covered by the reads produced by next-generation sequencing.

Given the biological and medical importance of CNVs, it is mandatory to develop probabilistic models that easily enable the detection CNVs from next-generation sequencing data. While other models have been developed to detect CNVs in a genomic scale (Sepúlveda et al.

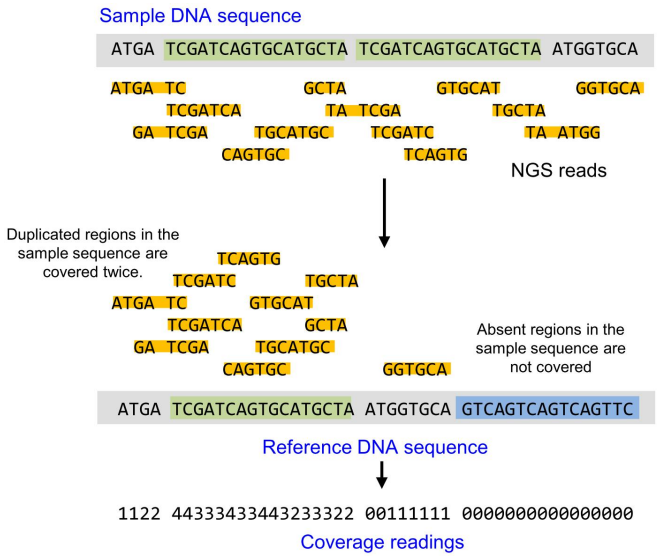


Figure 3.2. A schematic next-generation sequencing procedure to identify CNVs. Genomic DNA is shredded into fragments of manageable size. These fragments are partially sequenced as reads. Reads are subsequently aligned with a reference genome. The number of times a particular genomic site is covered by the reads constitute the coverage readings.

(2013) and Zhao et al. (2013) for a deep revision), an inferential approach that assists in the identification of CNVs in the exonic regions of genes, is in need. Exons are the regions of protein-coding genes that contribute to the protein (i.e. which are transcribed and translated into proteins; figure 3.1).

3.2 Objectives

The main objective of this thesis is to create an inferential model which permits to detect and identify CNVs on the exonic regions of genes using the next-generation sequencing coverage readings. In particular, it is intended:

- **Objective 1.** To present a suitable strategy to deal with the coverage noise (i.e. the coverage variations that are not related with the presence of exonic CNVs).
- **Objective 2.** To construct a probabilistic model for the exonic coverage ratio modeling.
- **Objective 3.** To create an inferential framework to detect the presence of CNVs in protein-coding genes, in a early phase of the CNVs analysis.

- **Objective 4.** To develop an inferential framework to identify CNVs, when a potential altered-gene is detected. The identification of CNVs includes determining the type of exonic CNVs (deletion, duplication, etc.).

3.3 Data structure

The analysis of CNVs is highly relative − what can be considered a duplication of a DNA fragment, can be considered a deletion in the fragment that was used to compare. Thus, the analysis of the exonic structure of genes may only be implemented using an exonic structure that we determine to be the standard. A standard exonic structure can be the most common in the analyzed population.

In addition, the comparison of coverage profiles indicate similar qualitative behaviors, suggesting they should be related up to a deformational constant (figure 3.3). Individual coverage readings can differ not only by the the presence of CNVs, but also by the reaction conditions of the next-generation sequencing (quantity of reagents and DNA in the sample, primer scheme, etc.), suggesting that some of the observed variability may be due to the experimental procedure.

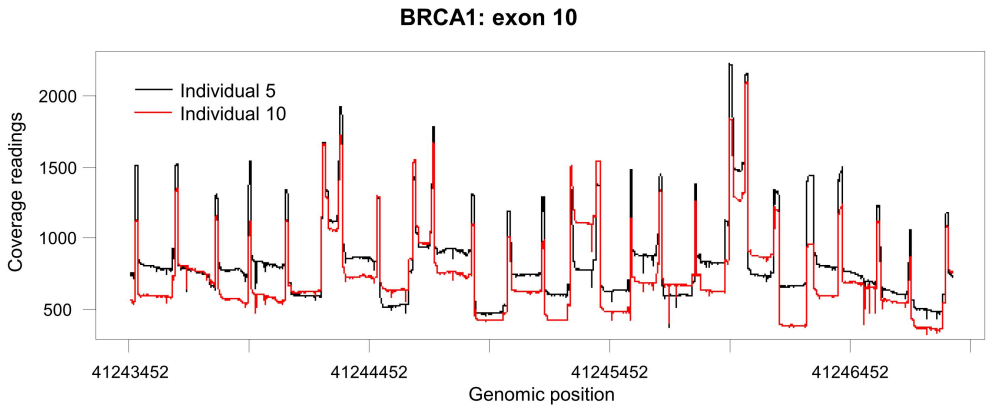


Figure 3.3. Comparison of two coverage profiles for the 10-th exon of the *BRCA*1 gene. The coverage readings peaks correspond to overlapping regions resulting from the next-generation sequencing and not to CNVs. As can be observed, these regions are common in both the coverage profiles.

It would be recommended to perform a data standardization which would minimize the effect of the next-generation sequencing reaction conditions, evidencing the unexpected features of the coverage readings. Consider that x_k is the coverage reading of the k-th genomic position. A longitudinal standardization of the individual coverage (expression 3.1) readings appears attractive but suffers from some problematic aspects.

$$\frac{x_k - \bar{x}_{..}}{s_{..}} \tag{3.1}$$

The longitudinal standardization can be highly biased by the noise introduced by the experimental variables. In addition, the longitudinal correction is highly sensitive to extreme observations, which can be the case when CNVs are present in the analysis. Thus, the estimated grouped coverage ratio mean and variance (biased by the CNV coverage readings) may dilute statistical signatures of CNVs. Finally, the longitudinal standardization do not allow multi-individual comparison analysis. CNV detection would be dependent on the mean and dimension of the case study individual coverage, being meaningless to compare statistics from different individuals.

3.3.1 Coverage ratio

We advance a standardization of the coverage readings that takes into account a reference coverage profile (equation 3.2). Consider that C_k^m is the observed coverage reading of the m-th individual in the k-th genomic position. Consider also the reference C^0 and the case study C^1 observed coverage readings, the coverage ratio is defined as:

$$y_k = \frac{C_k^1}{E[C_k^1|C_k^0]} = \frac{C_k^1}{f(C_k^0)} - 1 \tag{3.2}$$

$E[C_k^1|C_k^0]$ constitutes the expected case study individual coverage when a certain value of the reference coverage C_k^0 is observed. This factor eliminates the heterogeneity that depend on the experimental design/conditions and which are not associated with the presence of CNVs. In addition, the coverage ratio has a direct interpretation in the problem of CNVs detection and identification: exonic coverage ratios near to 0 would suggest the absence of a CNV, while coverage ratios near to 1 can be supportive of an additional copy (both conclusions were made taking in consideration the reference coverage). The main difficulty in calculating the coverage ratio is the determination of $E[C_k^1|C_k^0]$, which must be analyzed and computed for each gene in study.

Chapter 4

Hierarchical Bayesian model

Consider the CNV detection problem of comparing the means of the exonic coverage ratios θ_j of a group of J exons. Exons are discrete categories of a particular gene, in which we observed a random sample of n_j values of y_{ij} coverage ratios (figure 4.1). Chapter 4 presents and discusses the Bayesian approach that is used to model the CNV detection and identification problem.

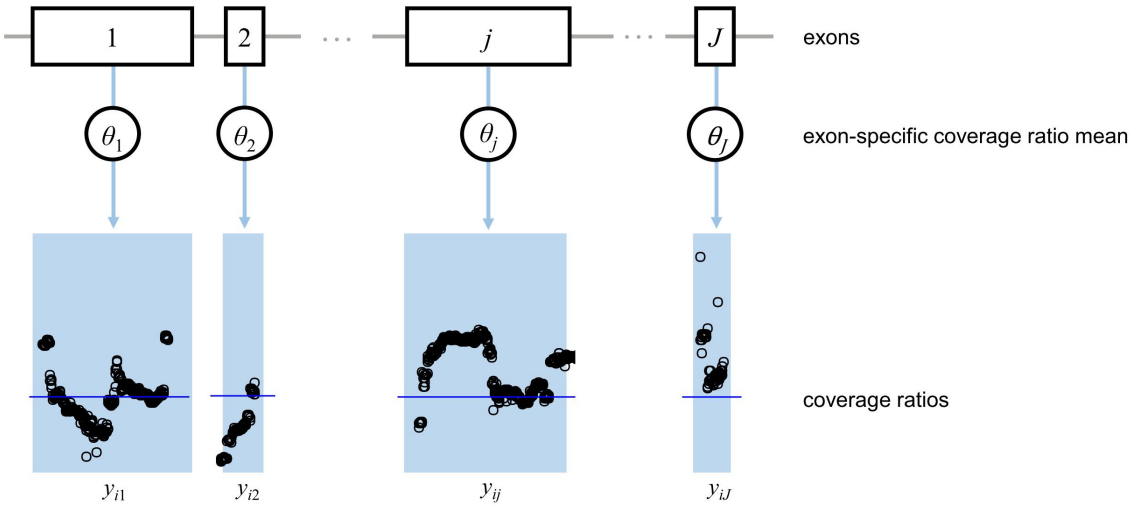


Figure 4.1. Schematic representation of the statistical problem underlying the detection and identification of CNVs.

4.1 Classical solution

The CNV detection problem can be formalized as:

$$\theta_1 = \theta_2 = ... = \theta_J$$

The statistical problem consists in compare the differences among J group means and can be solved using a classical analysis of variance (ANOVA). The idea behind ANOVA is simple: if the J group means appear significantly variable we chose separate sample means ($\hat{\theta_j} = \bar{y}_{.j}$); alternatively, if the variance between the group means is not significantly greater than the exon-specific variability, we should used the pooled mean estimate ($\hat{\theta} = \bar{y}_{..}$).

ANOVA methodology presents some assumption that makes it fragile in applications: one is the need of many observations per group and the other is its philosophy. ANOVA establishes, as its name implies, not a comparison of means but an analysis of variance. Therefore, even if ANOVA was useful determining the existence of at least one different exonic coverage ratio mean, later identification would be difficult. We could implement the Tuckey tests, which perform paired comparisons of the exon means, but this analysis requires a strict p-value adjustment and do not allow CNV-type characterization. Moreover, it is very common to find logic inconsistencies in the Tuckey tests.

We could also use, alternatively, a non-parametric approach, but the difficulty of identify CNVs still remains. Classical methods have all in common a set of limitative aspects where the Bayesian framework is a clear advantage:

- both gene and exon-level inferences must be integrated for the detection and identification phases of CNVs
- the exon inferences are meant to be individual-dependent (and not populational), as it is proper of diagnosis analysis
- the parameters uncertainty must be used to performed inference and their accuracy and precision should be accounted to support conclusions.

4.2 Hierarchical model

The problem of the CNV detection involves multiple parameters (θ and possibly their variances) which represent descriptive parameters of the same genomic unit, the gene, thus implying that a joint probability model should reflect they dependence.

Naturally, such a problem should be modeled hierarchically, with the coverage ratios modeled conditionally on certain parameters, known as hyperparameters: ϕ. The exonic coverage ratio means θ_j are considered to be random samples from the hyperparameters distribution.

4.2.1 Exchangeability in hierarchical models

If no other information exists to distinguish any of the θ_j apart from the data y, and no ordering or grouping of the parameters can be made, we must assume symmetry among the parameters in their prior distribution (Gelman et al., 2014f). The symmetry is represented probabilistically by exchangeability, being a necessary condition to build hierarchical models. By definition, the coverage ratio means are exchangeable in their joint distribution if $p(\theta_1, ..., \theta_J)$ is invariant to permutations of indexes $(1, ..., J)$ (Bernardo, 1996). In practice, ignorance implies exchangeability and we do not have reasons to believed that a particular exonic θ_j would have differentiated coverage ratio mean − we have no information to distinguish or group the J exons.

The simplest form of an exchangeable distribution considers each of the θ_j as independent samples from a prior distribution governed by the same unknown hyperparameter vector ϕ (Gelman et al., 2014f),

$$p(\theta|\phi) = \prod_j p(\theta_j|\phi)$$

4.2.2 Hyperparameters

The hyperparameters (ϕ) distribution is not known and thus has its own prior distribution $p(\phi)$. The joint prior distribution is

$$p(\phi, \theta) = p(\phi)p(\theta|\phi)$$

and the joint posterior distribution is

$$\begin{aligned} p(\phi, \theta|y) &\propto p(\phi, \theta)p(y|\theta, \phi) \\ &= p(\phi)p(\theta|\phi)p(y|\theta) \end{aligned}$$

with the latter simplification holding because the hyperparameters ϕ affect y only though θ and thus, the likelihood $p(y|\phi, \theta)$ depends only on θ (Gelman et al., 2014f). In order to create a joint probability model for (θ, ϕ) we must assign a prior distribution to ϕ. Usually, if little is known about ϕ we can assign a diffuse prior distribution, choice that will require to check that the resulting posterior distribution is proper (Gelman et al., 2014f).

4.3 Robust hierarchical model

To model the exonic coverage ratios θ_j, we implemented a robust hierarchical Bayesian model. The robust hierarchical approach assigns a t-student distribution (Gelman et al., 2014d), which is a heavy-tailed distribution and allows for extra dispersion in the data (a common limitation of the standard normal models). Before considering the probabilistic aspects of the model, let us first introduce the model parameters in the context of CNV detection and identification problem (figure 4.2). μ represents the mean of exonic coverage ratio means. τ is the scale of the exonic coverage ratio means. μ and τ can be useful parameters, informing whether a CNV is likely to exist in the group of exons, characterizing the gene-specific behavior of the coverage ratio. θ_j represent the j-th exon coverage ratio mean and V_j is the j-th exon coverage ratio variance. These parameters are as many as the number of exons in the case study gene and characterize the exon-specific behavior of the coverage ratio, being useful determining the probability of a specific exon to possess a CNV.

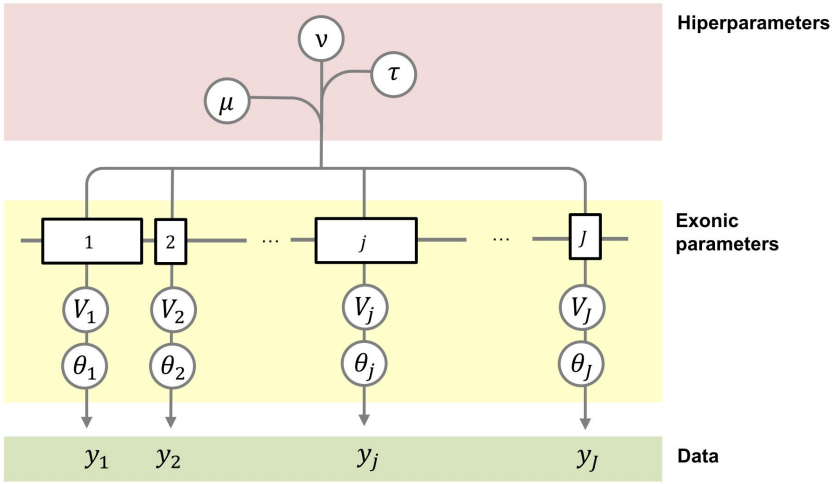


Figure 4.2. Structure of the hierarchical model for the CNV detection and identification problem.

We pretend to leave our model as general as possible for any gene we might want to study, thus we expect to accommodate well behaved coverage ratios but also, occasional extreme observations, which the presence of CNVs would probably promote. We opt for a robust t-student hierarchical approach to model the exon-specific coverage means: $p(\theta) = t_\nu(\mu, \tau^2)$.

$$p(\theta) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\tau}\left[1 + \frac{1}{\nu}\left(\frac{\theta-\mu}{\tau}\right)^2\right]^{-\frac{\nu+1}{2}}$$

The t distribution is characterized by a center μ, a scale τ and a degree of freedom parameter ν. ν is not an interpretable parameter on the biological problem of the CNV detection and

identification, however it can be very useful controlling the tail behavior of the t-distribution, which is, the propensity for outlier events in the group of the exonic coverage ratio means. The quantitative behavior of ν can accommodate a diverse fashion of possible probabilistic distributions for θ: $\nu = 1$ represents the particular case of a Cauchy distribution and as $\nu \to \infty$, the t-student distribution approaches the normal distribution, which is simply the normal hierarchical model.

Consider J independent exons, with exon j estimating the parameters θ_j and V_j from n_j independent data points, the j-th exon coverage ratio values y_{ij}. In order to perform Bayesian inference on the hierarchical model we must determine their posterior full probability model.

$$p(\theta, V, \mu, \tau, \nu|y) \propto p(\theta, V, \mu, \tau, \nu)p(y|\theta, V, \mu, \tau, \nu) \tag{4.1}$$

4.3.1 Prior distribution

The prior distribution must be set for the model parameters and hyperparameters. The joint probability distribution can be usefully decomposed in the three terms.

$$p(\theta, V, \mu, \tau, \nu) \propto p(\theta|V, \mu, \tau, \nu)p(V|\mu, \tau, \nu)p(\mu, \tau, \nu)$$

For the hyperparameters a uniform prior is assigned. Indeed, none or almost none information about μ, τ and ν are available prior the realization of the coverage measurements that would worth an informative prior. In addition, the hyperparameters are not observed in the experience, which compromise advancing distributional considerations.

The assumption behind the t-model, which considers that $\theta_j|\nu, \tau, \mu$ follows a $t_\nu(\mu, \tau^2)$ distribution can be alternatively thought as a mixture of normal distributions with a common mean and variances following a scaled inverse-χ^2 distribution. Thus, the conditional distributions of the t-student model define the priors of the exonic parameters.

$$\theta \mid V, \mu \sim N(\mu, V) \tag{4.2}$$

$$V \mid \tau, \nu \sim \text{Inv}{-}\chi^2\left(\nu, \tau^2\right) \tag{4.3}$$

A scaled inverse-χ^2 distribution is a special case of the inverse gamma distribution with shape and scale parameters equals to $\frac{\nu}{2}$ and $\frac{\nu}{2}\tau^2$ respectively. If V follows a scaled inverse-$\chi^2(\nu, \tau^2)$ then,

$$p(V) = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\tau^\nu V^{-\left(\frac{\nu}{2}+1\right)} \exp\left\{-\frac{\nu\tau^2}{2V}\right\} \qquad V > 0$$

To obtain a simulation draw from a scaled inverse-$\chi^2(\nu, \tau^2)$ distribution, we first draw Y from the χ^2_ν and then we compute $X = \frac{\nu\tau^2}{Y}$. This property is useful for computational implementations.

4.3.2 Likelihood

The coverage ratio are considered to be normally distributed, and thus the likelihood of each i-th observation in the j-th exon, y_{ij} follows $\mathrm{N}(\theta_j, V_j)$.

$$\begin{aligned} p(y_{.j}|\theta_j, V_j) &\propto V_j^{-\frac{n_j}{2}} \exp\left\{-\frac{1}{2V_j}\sum_i (y_{ij}-\theta_j)^2\right\} \\ &\propto V_j^{-\frac{n_j}{2}} \exp\left\{-\frac{1}{2V_j}[(n_j-1)s_{.j}^2 + n_j(\bar{y}_{.j}-\theta_j)^2]\right\} \end{aligned}$$

The likelihood can be written in terms of the sufficient statistics $\bar{y}_{.j} = \frac{1}{n_j}\sum_i y_{ij}$ and $s_{.j}^2 = \frac{1}{n_j-1}\sum_i (y_{ij}-\bar{y}_{.j})^2$.

Since we have no prior information to distinguished exons, the between exon observations should be considered exchangeable. Thus, the total likelihood becomes the product of the J exon-specific normal densities.

$$p(y|\theta, V) = \prod_j p\left(y_{.j}|\theta_j, V_j\right) \tag{4.4}$$

4.3.3 Posterior distribution

Using the prior and the likelihood (equations 4.2 to 4.4) we obtain the posterior distribution using the Bayes rule (equation 4.1).

$$\begin{aligned} p(\theta, V, \mu, \tau, \nu \mid y) &\propto N(\theta \mid \mu, V)\mathrm{Inv}{-}\chi^2\left(V \mid \nu, \tau\right)\prod_j p\left(y_{.j}|\theta_j, V_j\right) \\ &\propto \frac{\left(\frac{\nu}{2}\right)^{J\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)^J}\tau^{J\nu}\exp\left\{-\sum_j \frac{1}{2V_j}[\nu\tau^2 + (\theta_j-\mu)^2 + (n_j-1)s_{.j}^2 + n_j(\bar{y}_{.j}-\theta_j)^2]\right\} \\ &\qquad \prod_j \left[V_j^{-\left(-\frac{n_j+3+\nu}{2}\right)}\right] \end{aligned} \tag{4.5}$$

4.3.4 Posterior conditional distributions

The posterior distribution (equation 4.5) is a complex expression, however their full conditional posterior distributions are, except for ν, known cases.

- Conditional posterior distribution of μ
  Conditional on the data y and the other parameters of the model, information about μ is supplied by θ_j, each with its own variance V_j. Combining with the uniform prior distribution on μ yields,

$$\begin{aligned} p(\mu|.,y) &\propto \exp\left\{-\frac{1}{2}\sum_j \frac{(\theta_j-\mu)^2}{V_j}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\sum_j \frac{1}{V_j}\right)\left(\mu - \frac{\sum_j \frac{\theta_j}{V_j}}{\sum_j \frac{1}{V_j}}\right)^2\right\} \\ &= N\left(\frac{\sum_j \frac{1}{V_j}\theta_j}{\sum_j \frac{1}{V_j}}, \frac{1}{\sum_j \frac{1}{V_j}}\right) \end{aligned} \tag{4.6}$$

- Conditional posterior distribution of τ^2
  Conditionally on the data y and the other parameters of the model, all the information about τ comes from the variances V_j and ν. The posterior is a gamma distribution.

$$\begin{aligned} p(\tau^2|.,y) &\propto \tau^{J\nu}\exp\left\{-\frac{\nu\tau^2}{2}\sum_j \frac{1}{V_j}\right\} \\ &= \text{Gamma}\left(\frac{J\nu}{2}+1, \frac{\nu}{2}\sum_j \frac{1}{V_j}\right) \end{aligned} \tag{4.7}$$

- Conditional posterior distribution of θ_j
  In the joint posterior density the factors that involve θ_j are the $N(\mu,\tau^2)$ prior distribution and the normal likelihood. Conditional on the hyperparameters and the vector of exon variances V, information about θ_j is supplied by μ and the n_j coverage ratio observations y_{ij}.

$$\begin{aligned} p(\theta_j|.,y) &\propto \exp\left\{-\frac{1}{2V_j}[(\theta_j-\mu)^2 + n_j(\bar{y}_{.j}-\theta_j)^2]\right\} \\ &\propto \exp\left\{-\frac{n_j+1}{2V_j}\left(\theta_j - \frac{\mu+\bar{y}_{.j}n_j}{n_j+1}\right)^2\right\} \\ &= N\left(\frac{\mu+\bar{y}_{.j}n_j}{n_j+1}, \frac{V_j}{n_j+1}\right) \end{aligned} \tag{4.8}$$

- Conditional posterior distribution of V_j
  Conditional on the data y and the other parameters of the model and with a normal likelihood, each V_j has a scaled inverse-χ^2 posterior distribution.

$$\begin{aligned} p(V_j|.,y) &\propto V_j^{-\frac{n_j+3+\nu}{2}} \exp\left\{-\frac{1}{2V_j}\left[\nu\tau^2+(\theta_j-\mu)^2+(n_j-1)s_{.j}^2+n_j(\bar{y}_{.j}-\theta_j)^2\right]\right\} \\ &= \text{Inv}{-}\chi^2\left(n_j+1+\nu, \frac{\nu\tau^2+(n_j-1)s_{.j}^2+(\theta_j-\mu)^2+n_j(\bar{y}_{.j}-\theta_j)^2}{n_j+1+\nu}\right) \\ &= \frac{\nu\tau^2+(\theta_j-\mu)^2+(n_j-1)s_{.j}^2+n_j(\bar{y}_{.j}-\theta_j)^2}{\chi^2_{n_j+1+\nu}} \end{aligned} \tag{4.9}$$

- Conditional posterior distribution of ν
  The conditional distribution of ν does not take a simple form.

$$p(\nu|.,y) \propto \frac{\left(\frac{\nu}{2}\right)^{J\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)^J}\tau^{J\nu}\exp\left\{-\frac{\nu\tau^2}{2}\sum_j\frac{1}{V_j}\right\}\prod_j\left[V_j^{-\left(-\frac{n_j+3+\nu}{2}\right)}\right] \tag{4.10}$$

4.3.5 Gradient Vector

For the Hamiltonian Monte Carlo algorithm we need the gradient of the log posterior density (table 4.1). Within the robust hierarchical model the operations are easily performed. Let $\lambda = (\mu, \tau, \theta, V, \nu)$.

$$\begin{aligned} \log p(\lambda|y) =& \frac{J\nu}{2}\log(\nu/2) - J\log\Gamma(\nu/2) + J\nu\log\tau \\ &- \sum_j \frac{1}{2V_j}[\nu\tau^2+(\theta_j-\mu)^2+(n_j-1)s_{.j}^2+n_j(\bar{y}_j-\theta_j)^2] \\ &- \frac{1}{2}\sum_j(n_j+3+\nu)\log V_j \end{aligned}$$

An Hamiltonian Monte Carlo algorithm with constrained parameters can lead the trajectory outside the boundary, thus wasting some iterations. We have three positive real parameters: τ, V_j and ν. One remedy is to transform the space to be unconstrained. In this case, the simplest way to handle the constraint $\tau > 0$ is to transform to $\log\tau$.

This requires that the posterior must be multiplied by the Jacobian τ, which means adding the $\log\tau$ to the log posterior. The gradient also changes, considering the adding term on the log posterior and the necessary subsequent multiplication by the Jacobian. We proceed to the V_j and ν log-transformation as we explained for τ.

Parameter	Component	Expression
μ	$\frac{\partial \log p(\lambda\|y)}{\partial \mu}$	$\sum_j \frac{1}{V_j}(\theta_j - \mu)$
τ	$\frac{\partial \log p(\lambda\|y)}{\partial \log \tau}$	$J\nu - \nu\tau^2 \sum_j \frac{1}{V_j} + 1$
θ_j	$\frac{\partial \log p(\lambda\|y)}{\partial \theta_j}$	$\frac{1}{V_j}[\mu + n_j\bar{y}_{.j} - (1+n_j)\theta_j]$
V_j	$\frac{\partial \log p(\lambda\|y)}{\partial \log V_j}$	$\frac{1}{2V_j}[\nu\tau^2 + (\theta_j - \mu)^2 + (n_j - 1)s^2_{.j} + n_j(\bar{y}_j - \theta_j)^2] - \frac{n_j+3+\nu}{2} + 1$
ν	$\frac{\partial \log p(\lambda\|y)}{\partial \log \nu}$	$\frac{J\nu}{2}(\log\nu/2 + 1) - \frac{J\nu}{2}\frac{\log\Gamma(\nu/2)}{d\nu/2} + J\nu\log\tau - \frac{\nu}{2}\sum_j \log V_j - \frac{\nu\tau^2}{2}\sum_j \frac{1}{V_j} + 1$

Table 4.1. The gradient vector components for the Hamiltonian Monte Carlo algorithm. The τ, V and ν parameters were log-transformed.

Considering the log-transformations of the τ, V and ν parameters, the position vector for the Hamiltonian Monte Carlo becomes $\lambda = (\mu, \log\tau, \theta, \log V, \log\nu)$ and the log posterior distribution,

$$\begin{aligned}\log p(\lambda|y) =& \frac{J\nu}{2}\log(\nu/2) - J\log\Gamma(\nu/2) + J\nu\log\tau \\ & - \frac{1}{2}\sum_j \frac{1}{V_j}[\nu\tau^2 + (\theta_j - \mu)^2 + (n_j - 1)s^2_{.j} + n_j(\bar{y}_j - \theta_j)^2] \\ & \frac{1}{2}\sum_j (n_j + 3 + \nu)\log V_j \\ & + \log\tau + \sum_j \log V_j + \log\nu\end{aligned}$$

4.4 Computational implementation

The hierarchical robust model developed here was implemented in the R language. In a first phase the algorithm extracts the next-generation sequencing data for the desired genomic coordinates. The files with the coverage readings are of the `.bam` type and the genomic coordinates are of the `.bai` type. To extract the coverage readings the R package `rbamtools` was used (Li et al., 2009).

After the extraction of the data it is implemented the Hierarchical Bayes. Taking advantage of the marginal posteriors of each parameter it was constructed an iterative MCMC scheme. For those parameters which possess well know posterior marginal distributions (i.e. μ, τ^2, θ and V, equations 4.6 to 4.9) the Gibbs sampler was used. Because the parameter ν has a complex and unknown posterior marginal (equation 4.10) we have used, alternatively, the Metropolis-Hastings algorithm. The Gibbs and the Metropolis Hastings techniques were combined in building-blocks, that proceed iteratively. A schematic outline of the computational implementation and the R code are represented in the Appendix A.

```
rnu <- function(nu,tau,V,snu) {
  inu <- rnorm(1,1/nu,snu)
  if (inu <= 0 | inu >1) {
    jp <- 0
  } else {
    nus <- 1/inu
    r   <- exp( nus*( J*log(nus/2)/2 + J*log(tau) -
                sum(log(V))/2 - tau*tau*sum(1/V)/2 ) - J*lgamma(nus/2) -
                nu*( J*log(nu /2)/2 + J*log(tau) -
                sum(log(V))/2 - tau*tau*sum(1/V)/2 ) + J*lgamma(nu /2)  )
    if (runif(1) < r) {
      nu <- nus
    }
    jp <- min(r,1)
  }
  return(c(nu,jp))
}
```

Figure 4.3. Metropolis-Hastings algorithm implemented in R for simulating the ν parameter on the Hierarchical Bayesian robust model.

4.4.1 Metropolis-Hastings step for $1/\nu$

Simulating $1/\nu$ using the Metropolis-Hastings algorithm requires the definition of a proposal distribution. We used the normal density, with mean in the last sampled iterate and standard deviation `snu` (figure 4.3).

The standard deviation or the dimension parameter of the proposal distribution must be defined carefully. In one hand, it must be considered that lower values of `snu` promote intra-chain correlation, leading to the undesirable random walk behavior of ν iterates (and also on the dependent parameters, as τ and V). In the other hand, if `snu` is too high can lead to higher rates of rejection which decrease the efficiency of the algorithm. We proceed with some simulations and conclude that `snu` between 0.10 and 0.08 guarantees a mean acceptance probability of approximately 44%, the recommended acceptance probability for one dimension (Gelman et al., 2014b).

4.4.2 Hamiltonian Monte Carlo

To avoid the slow exploration of the parameters space in the posterior distribution (as were common to the ν, τ^2 and V parameters) we implemented the Hamiltonian Monte Carlo algorithm in conjunction with the Gibbs and the Metropolis Hastings building-blocks.

```
hmc <- function(lambda,M,epson,mj,s2j,nj){
  L  <- floor(1/epson)-1
  D  <- length(lambda)

  #RANDOM MOMENTUM
  psi     <- rnorm(D,0,sqrt(M))
  lambdai <- lambda
  lposti  <- lpost(lambda,mj,s2j,nj) - sum(psi*psi/M)/2

  #LEAPFROG STEP
  psi    <- psi + epson*lgrad(lambda,mj,s2j,nj)/2
  for (i in 2:(L-1)) {
    lambda <- lambda + epson*psi/M
    psi    <- psi + epson*lgrad(lambda,mj,s2j,nj)
  }
  psi <- psi + epson*lgrad(lambda,mj,s2j,nj)/2

  #METROPOLIS STEP
  r <- exp(lpost(lambda,mj,s2j,nj) - sum(psi*psi/M)/2 - lposti)
  if (is.na(r) == T) { r <- 0 }
  if ( runif(1) <= min(1,r) ) {
    return(c(1,lambda))
  } else {
    return(0)
  }
}
```

Figure 4.4. Hamiltonian Monte Carlo algorithm implemented in R. `lgrad` and `lpost` correspond to the gradient and log-posterior functions respectively.

The Hamiltonian Monte Carlo algorithm uses as input the ϵ value, the M mass matrix (figure 4.4) and the log-posterior gradient vector (table 4.1). The ϵ value is used to calculate the necessary leapfrog iterates, considering the condition $L\epsilon = 1$. The M mass matrix is defined by default as a diagonal matrix, with diagonal vector $m = \mathbf{1}$.

Chapter 5

Case study: *BRCA*1 and *BRCA*2 genes

Chapter 5 includes a discussed application of the robust hierarchical Bayesian model to detect and identify CNVs in the *BRCA*1 and *BRCA*2 genes. *BRCA*s are protein-coding-genes with a known exonic structures of 23 and 26 expressed exons (figure 5.1).

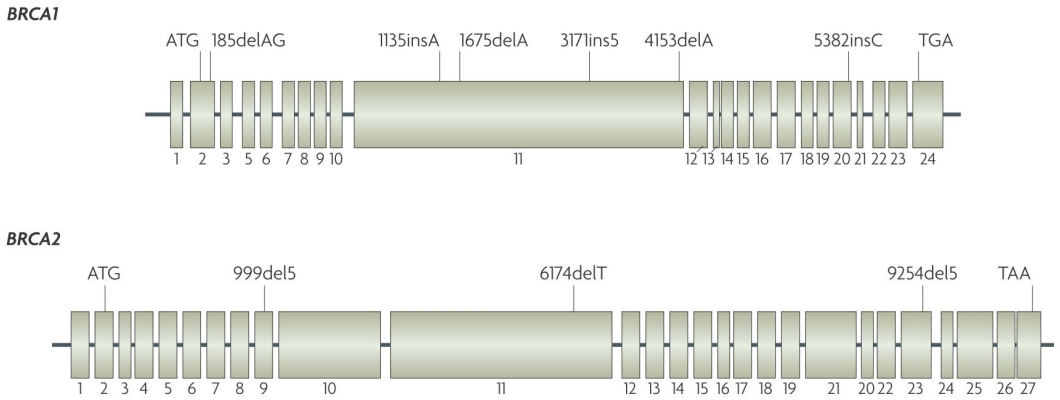


Figure 5.1. Schematic view of the *BRCA*1 and *BRCA*2 genes in the human genome. Exons are in relative sizes. Retrieved from Fackenthal and Olopade (2007).

5.1 *BRCA*1 and *BRCA*2 genes

*BRCA*1 and *BRCA*2 genes encode proteins playing a role on the repair of the genetic material. Mutation or structural alterations on these genes may compromise the proper repair of the altered DNA which frequently leads cells from growing and dividing too rapidly or in an uncontrolled way (Fackenthal and Olopade, 2007).

Mutations on *BRCA*1 and *BRCA*2 genes increase woman's risk to develop breast and ovarian cancer (Easton, 1999; Pal et al., 2005). Studies have shown that CNVs in the *BRCA*1 and *BRCA*2 genes could also lead to breast and ovarian cancer development (Krepischi et al., 2012; Kuusisto et al., 2013).

*BRCA*1 and *BRCA*2 genes are present in different regions of the Human genome (table 5.1). *BRCA*1 is located on the chromosome 17, in the 41 197 694 to 41 277 287 genomic coordinates and comprising 22 expressed exons. *BRCA*2 is located on the chromosome 13, between the 32 890 558 and 32 973 809 genomic coordinates and possess 26 expressed exons. In both genes the exon 1 is not expressed (table 5.1 and figure 5.1).

5.1.1 Experimental design

We analyze 10 individuals for the *BRCA*1 and *BRCA*2 genes by the next-generation sequencing technique under similar reaction conditions. The *BRCA*1 and *BRCA*2 exonic coverage readings were assessed for all the individuals, and the individual 10 was used as reference, to calculate the other individual's coverage ratio. Individual 10 was chosen for having a typical genetic structure for both the *BRCA*1 and *BRCA*2 genes. Individual 7 and 9 possess exonic CNVs determined by alternative methodologies:

- Individual 7, *BRCA*2, exon 20
- Individual 9, *BRCA*1, exon 16
- Individual 9, *BRCA*2, exons 19 and 20

Gene	Genomic coordinates	
BRCA1	41197694	41277287
exon 2	41276034	41276113
exon 3	41267743	41267796
exon 4	41258473	41258550
exon 5	41256885	41256973
exon 6	41256139	41256278
exon 7	41251792	41251897
exon 8	41249261	41249306
exon 9	41247863	41247939
exon 10	41243452	41246877
exon 11	41242961	41243049
exon 12	41234421	41234592
exon 13	41228505	41228631
exon 14	41226348	41226538
exon 15	41222945	41223255
exon 16	41219625	41219712
exon 17	41215891	41215968
exon 18	41215350	41215390
exon 19	41209069	41209152
exon 20	41203080	41203134
exon 21	41201138	41201211
exon 22	41199660	41199720
exon 23	41197695	41197819

Gene	Genomic coordinates	
BRCA2	32890558	32973809
exon 2	32890598	32890664
exon 3	32893214	32893462
exon 4	32899213	32899321
exon 5	32900238	32900287
exon 6	32900379	32900419
exon 7	32900636	32900750
exon 8	32903580	32903629
exon 9	32905056	32905167
exon 10	32906409	32907524
exon 11	32910402	32915333
exon 12	32918695	32918790
exon 13	32920964	32921033
exon 14	32928998	32929425
exon 15	32930565	32930746
exon 16	32931879	32932066
exon 17	32936660	32936830
exon 18	32937316	32937670
exon 19	32944539	32944694
exon 20	32945093	32945237
exon 21	32950807	32950928
exon 22	32953454	32953652
exon 23	32953887	32954050
exon 24	32954144	32954282
exon 25	32968826	32969070
exon 26	32971035	32971181
exon 27	32972299	32972907

Table 5.1. Genomic coordinates for *BRCA*1 and *BRCA*2 genes exonic regions.

5.2 Reference coverage transformation

We propose the use of a reference coverage readings to standardize the case study individual coverage readings, using the expected case study coverage given the standard coverage: $f(C^0) = E[C^1|C^0]$ (equation 3.2). We firstly analyzed the type of relationship between the coverages, and we found it clearly linear for both the *BRCA*1 and *BRCA*2 genes (figure 5.2). In addition, we consider that the genomic regions uncovered by the experimental procedure should return 0 coverage in both, the case study and the reference individuals, thus implying that the linear relationship between coverages should be of the type $f(C^0) = aC^0$.

We used the minimum mean square error estimator to calculate the value of a. We obtained,

$$\hat{a} = \frac{\sum_i C_i^0 C_i^1}{\sum_i (C_i^0)^2}$$

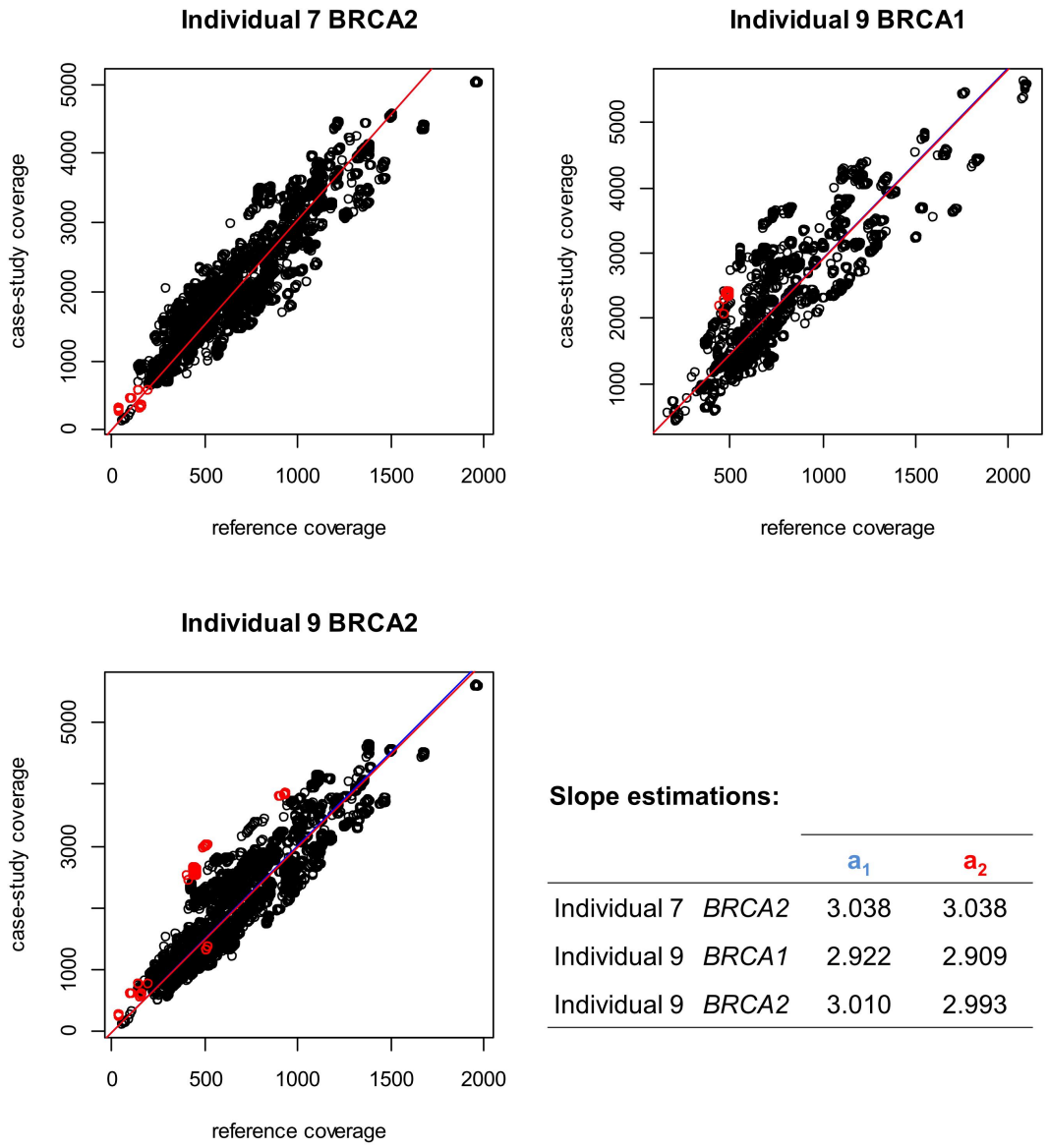
In normal applications we do not know whether one or more CNVs are present in the case study individual, which naturally, could interfere in the computation of $\hat{a}$. During the standardization phase, we would want the potential CNV coverage readings interfere as little as possible, so they can be more easily identified later, using a proper statistical analysis. Using the coverage readings from the individuals which possess exonic CNVs (individuals 7 and 9), we compare the estimates of $\hat{a}$ with and without the coverage points corresponding to the altered-exons (figure 5.2). We conclude that the presence of coverage readings from exonic CNVs were of minimal importance when computing $\hat{a}$, and by consequence, when computing the coverage ratios.

We proceed thus without performing any correction to a.

5.3 MCMC output analysis

We proceed with the MCMC algorithm for 2 chains and 30 000 iterations. We thin the resulting chain keeping every 20th iterate, thus 1500 iterates were kept for further analysis. The first 500 iterates were burned-in.

The convergence and mixing of the MCMC iterates was evaluated using both, the potential scale reduction index ($\hat{R}$) and visual inspection (Gelman and Rubin, 1992). The $\hat{R}$ was estimated for all the parameters of the model and for each iterate, but the $\max\{\hat{R}\}$ was assessed to globally determined the convergence and mixing of each chain. The condition $\max\{\hat{R}\} < 1.01$ was used and verified for all runs. Visual inspection of the MCMC chains was employed via trace plots (parameter draw *vs.* iterate number) and autocorrelation plots to additionally validate convergence (within chain behavior), mixing (between chain behavior) and independence of the iterates.



Slope estimations:

		a_1	a_2
Individual 7	*BRCA2*	3.038	3.038
Individual 9	*BRCA1*	2.922	2.909
Individual 9	*BRCA2*	3.010	2.993

Figure 5.2. Comparison of the linear estimates of $E[C^1|C^0]$, the expected case study coverage given the standard coverage, by accounting (blue line, slope a_1) and not accounting (red line, slope a_2) with the exonic CNVs.

The converged, well mixed and independent MCMC iterates constitute random draws from the posterior distribution and were used to estimate the posterior model parameters. Here we have used the quadratic loss criteria (i.e. the mean) to provide an point estimate of the model parameters.

5.4 Model assessment

The model assessment is necessary to validate the parametric assumptions that were made when the hierarchical robust model was constructed:

- The coverage ratio is normally distributed: $y_{ij} \sim \mathrm{N}(\theta_j, V_j)$
- The coverage ratio means are t-student distributed: $\theta \sim t_\nu(\mu, \tau^2)$
- A uniform prior distribution was assigned to the hyperparameters.

5.4.1 Sensitivity analysis

The hyperparameters of the hierarchical Bayesian robust model were considered as having a uniform prior distribution: $p(\mu, \tau^2, 1/\nu) \propto 1$. The uniform prior is a non-informative prior and it is particularly suited for this case study as neither we have previous information on the hyperparameters, nor they are directly observed during the experience to validate a potential informative prior.

It is thus essential to analyze the marginal posterior distributions of the hyperparameters in order to check if the data information was properly captured. We conclude that the marginal posteriors of the hyperparameters do give reasonable posterior information (figure 5.3): they all posses non-flat distributional features with a clear mode. Similar results were observed in others individuals/genes.

Other aspect to take in consideration is the propriety of the posterior distribution. The use of a uniform prior distribution on the parameters with infinite support (μ and τ) leads to improper priors. However, as we already have shown (equations 4.6 and 4.7) the conditional posterior distributions of μ and τ are proper.

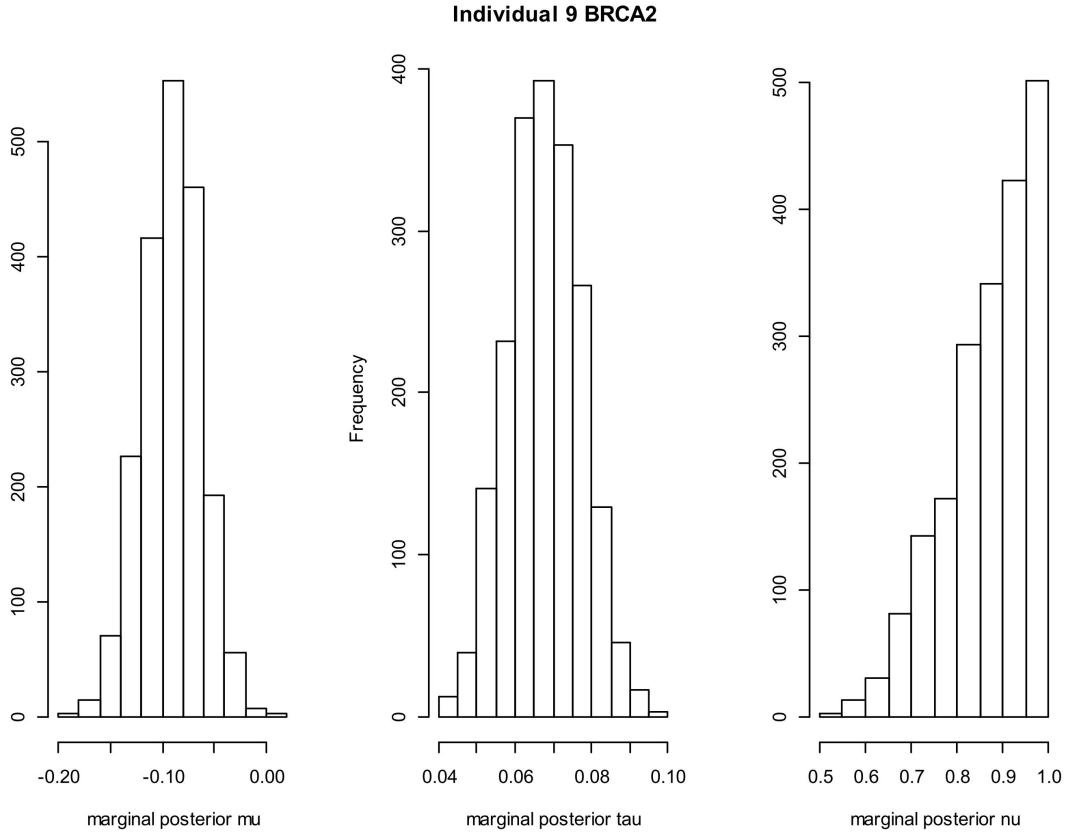


Figure 5.3. Histograms of the marginal posterior distributions of the hyperparameters μ, τ and $1/\nu$, based on converged, mixed and independent MCMC draws. This is a particular case were the presence of CNVs have been reported in the 19th and 20th exons.

5.4.2 Normal likelihood

The likelihood was assumed to be normal distributed: $y_{ij} \sim \mathrm{N}(\theta_j, V_j)$. To verify this assumption we proceed with the normalization of the coverage ratios

$$z_{ij} = \frac{y_{ij} - \theta_j}{\sqrt{V_j}}$$

considering the posterior estimates of the hierarchical robust model parameters. The normalized observations were used to calculate four summary statistics: mean, standard deviation, skewness and kurtosis. The observed summary statistics were compared with the values we would expect to obtain on a standard normal distribution.

The analysis of the expected and the observed summary statistics lead to the conclusion that the observed data deviates from the normal distributed data only for the third and forth order moments (figure 5.4). The model parameters (θ and V) satisfactory describe the mean and the variability of the exonic coverage ratios. In some particular cases, the variance of the normalized coverage ratios is out of the expected 95% intervals.

$$V[z_{ij}] = V\left[\frac{y_{ij} - \theta_j}{\sqrt{V_j}}\right] = \frac{V[y_{ij}]}{V_j}$$

For those exons we verify that $V[z_{ij}] < 1$, from which we conclude that $V[y_{ij}] < V_j$, suggesting that some V_j are being overestimated. For this particular analysis we have used the quadratic loss criteria (i.e. the mean) to estimate V. However, even with the median and mode estimates, the estimation of V maintains. This had to be checked since the posterior conditional of V (differently from θ) is skewed (equation 4.9). Notice that similar results were obtained for the other individuals/genes.

5.4.3 The robust model

Another aspect of the model that must be considered is the distributional assumption of the mean of the exonic coverage ratios: $\theta \sim t_\nu(\mu, \tau^2)$. While a parameter of centrality and variability (μ and τ) would be evident to govern the exonic coverage ratios, the use of an additional parameter ν of robustness/extremeness, with no biological interpretability in the context of the CNV detection and identification problem, can be arguable. The importance of ν can be better evaluated considering its effect on θ and V.

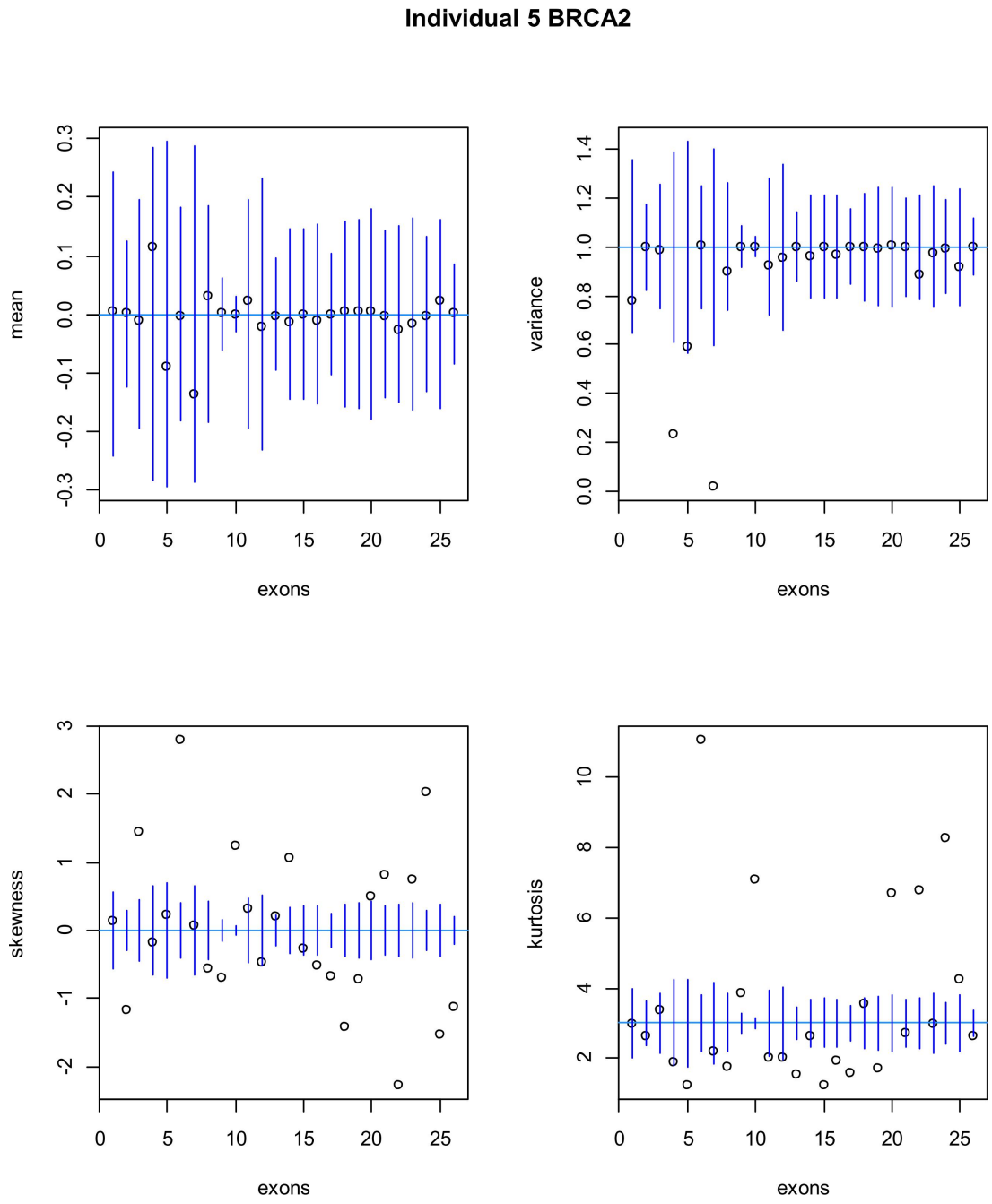


Figure 5.4. Analysis of four summary statistics (mean, standard deviation, skewness and kurtosis) for the observed normalized coverage ratios $z_{ij} = \frac{y_{ij}-\theta_j}{\sqrt{V_j}}$, considering the quadratic loss posterior estimates of the model parameters. The blue vertical lines correspond to exon-specific 0.95 simulated intervals, considering the distribution of the corresponding summary statistics for a n_j random draw of a standard normal distribution.

Using the converged, mixed and independent MCMC random draws of the posterior distribution, we analyze the joint behavior of ν and θ and V model parameters. The central estimations of θ and V do not significantly change according to ν, however their precision is affected, being higher when $1/\nu \to 1$. In addition, we must refer that the support of the posterior marginal of $1/\nu$ is clearly in the robust region of the model, having support between 0.6 and 1 in most of the analysis, which corresponds to a t-distribution with 1-1.6 degrees of freedom. Thus, the use of a robust model shows itself appropriate for the detection and identification of CNVs, particularly when considering the normal alternative (which would assume $1/\nu \to 0$).

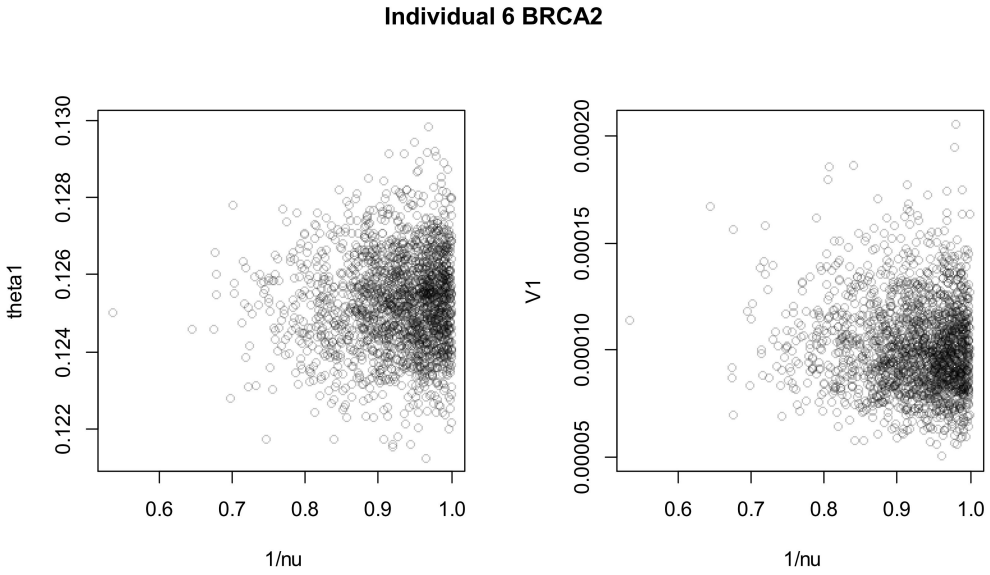


Figure 5.5. Joint probability distribution of ν with the model parameters θ and V for the individual 6 and *BRCA*2 gene. For sake of simplicity, the particular cases of the first element of the exonic coverage ration mean and variance vectors are shown (θ_1 and V_1).

5.5 Posterior predictive checking

In addition to the model assumptions, we may want to know the predictive quality of the model. The predictive posterior distribution $p(y^*|y)$ can be used to assess model predictability in a straightforward way. Lets consider a random draw of the posterior distribution. The values of θ and V can be used to replicate, by simulation, the observed values. The close the replicated values are from the observations, the higher is the model predictive quality.

The replicates and the observed values are generally compared using summary statistics of interest $T(.)$. Here we look closer to the mean and standard deviation. Once several replicated summary statistics are simulated, they can be compared with the corresponding

summary statistic, calculated with the observed data, and a Bayesian p-value can be assessed: $P[T(y^*) < T(y)]$ (Gelman et al., 2014g). The Bayesian p-values should approach 0.5 in good predictive conditions, while extreme values (near to 0 or 1), express poor predictive quality of the summary statistic (Gelman et al., 2014g).

Using the predictive distribution we found that the mean and the variance are generally well predicted by the model (figure 5.6). For some specific exons, we observed that the variance of the coverage ratio possessed Bayesian p-values near to 0. Considering that we are calculating $P[T(y^*) < T(y)] = 0$ then we conclude that $T(y^*) > T(y)$. This result is in congruence with the analysis we performed when checking for data normality, in which some exonic variances were overestimated.

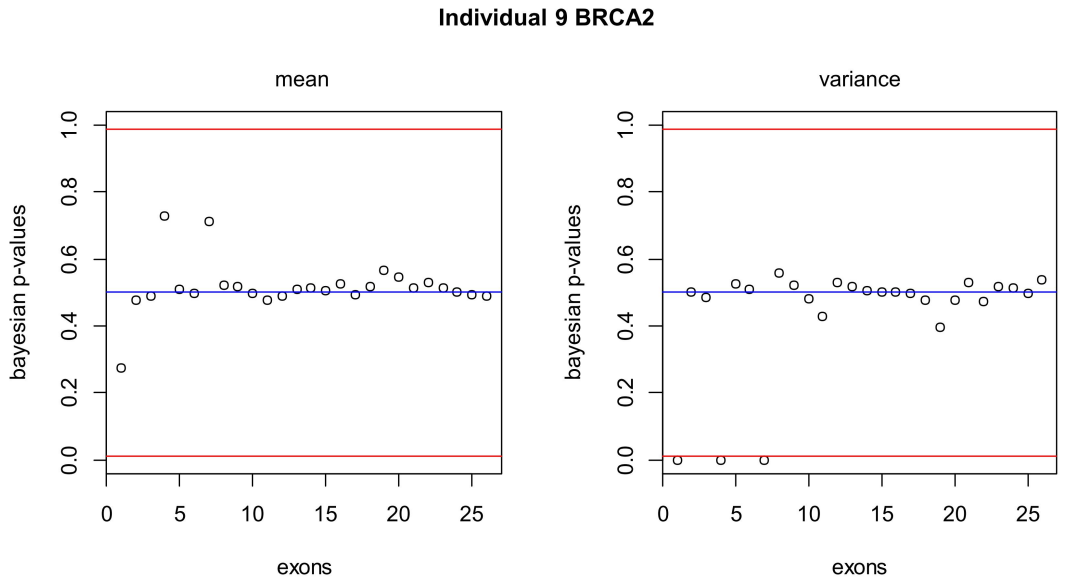


Figure 5.6. Model predictability analysis. The Bayesian p-values are shown for both, the mean (left plot) and the variance (right plot) summary statistics.

The predictive performance of the hierarchical robust model suggests it is able to produce accurate and precise inferences to detect and identify exonic CNVs. Some exonic variances showed to be overestimated, which is of minor importance, because we know that the coverage ratio means (i.e. the CNV-type) are being properly predicted and in addition, the overestimated variances avert the existence of false positives.

5.6 Inference of CNVs

Apart from the model assessment, another important characteristic of a model (if not the most important) is to give a proper answer to the problem it was meant to solve. In our case we pretend to know if the Bayesian robust model implemented here is capable of detect and identify CNVs in the *BRCA*1 and *BRCA*2 genes.

5.6.1 CNV detection

Hyperparameters can be used to detect CNVs, that is, to conclude whether a CNV is present in one or more exons of a particular gene, prior to the attempt to characterize the CNVs once they were discovered to exist. Most of the analysis will return negative results, and advance with an inferential strategy to rapidly indicate those that require a fine analysis is of clear interest.

The comparison of the hyperparameters estimates can provide some insight on the CNV detection (figure 5.7). μ do not show any special behavior in the presence of CNVs. $1/\nu$ and τ hyperparameters were expected to be more sensible to the presence of CNVs, as they are measures of the exonic coverage ratios variability. While the $1/\nu$ parameter is not particularly associated with the presence of CNVs, the τ parameter is. We checked the estimations of τ in the t-model for the analyzed individuals and found that in the three cases in which CNVs were detected, the scale parameter was higher (red arrows, figure 5.7). The definition of thresholds to perform the detection of CNVs would still require more sampling, particularly of CNV-positive cases, but τ can be appointed as a promising parameter.

5.6.2 CNV identification

The identification phase includes the characterization of exonic CNVs with particular interest in providing a criteria of statistical evidence. It is thus necessary to advance a strategy that not only characterize the type of structural variation we have observed (CNV-type), as well as, provide a quality parameter of such characterization, i.e. a measure of how much one can be sure about the presence of a CNV in a particular exon, considering a regular case. Here the regular case in the standard coverage.

CNVs are, by nature, discrete categories that must be considered in the identification process. Thus, we may want to discretize the log coverage ratios (a continuous variable) into meaningful intervals. Another aspect of the CNV identification that must be considered, is the possibility of its occurrence in heterozygosity. Humans receive two copies of each gene, and the duplicated (or the deleted) CNVs could have affected only one copy or both (case in

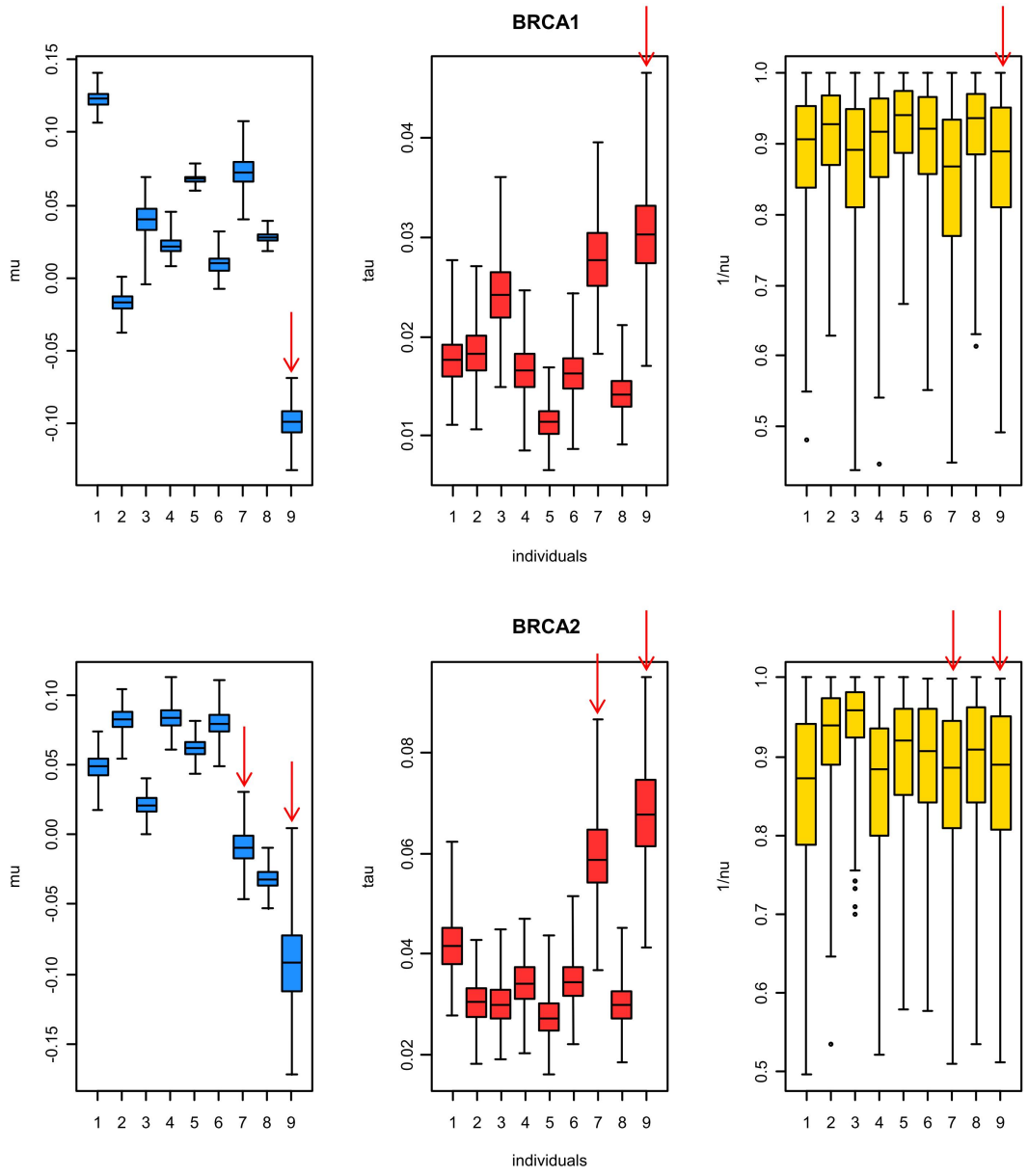


Figure 5.7. Representation of the model hyper parameters, μ (blue), τ (red) and ν (yellow) for each of the case study individuals and genes. Cases where CNVs have been reported are indicated with a red arrow.

which we would have a homozygous CNV). Consider the following example. A homozygous duplication (2-2) is likely to produced, in average, duplicated coverage ratios, however, in an heterozygous duplication (1-2) the coverage ratio would be, in average 1.5. Note that an exonic duplication and deletion in heterozygosity (0-2) would be confounded with a normal individual, at least considering the θ_j estimate (V_j would be most certainty higher). It must be stressed that these confounding scenarios are not a limitation of the model, but of the experimental technique.

Notice that we used a standardize coverage that considers only the deviance from the expected behavior (1) we should be aware that our normal region is between the gain in heterozygorizty (0.5) and the loss in heterozygozity (-0.5).

The visual inspection of θ and V parameters appears informative for the CNV identification (figure 5.8). The θ estimates in the individuals 7 and 9 for the $BRCA1$ and $BRCA2$ genes are clearly suggesting the presence of CNVs, because the exon-specific θ_j values lied out of the normal region of the coverage ratio (i.e. $0.5 <$ coverage ratio < -0.5). The non-altered (or neutral) region corresponds to the interval where the coverage readings of the case study individual follows the same pattern as the standard individual (here, the individual 10).

Apart from the visual inspection, we would want an inferential measure to validate the identification of potential CNVs. Here we follow the Bayesian hypothesis testing and advance with a Bayes factor approach on the predictive distribution. The Bayes factors were calculated considering the probability of the predicted exonic coverage ratio lies out of the normal region (homozygous for 1 copy).

$$BF = \frac{p(y^*|y > 0.5) + p(y^*|y < -05)}{p(-0.5 < y^*|y < 0.5)} \tag{5.1}$$

The predictive Bayes factors are calculated favoring the existence of CNVs.

Exon-specific Bayes factors can be used to determined the existence of a CNV in a particular exon if the j-th Bayes factor is higher than 1. Table 5.2 depicts some Bayes factors higher than 1, which correspond to those exons that possess CNVs (in bold). A closer analysis of the table is clear to show that artifact CNVs possess Bayes factors that are always less than 1, while true exonic-CNVs always possess Bayes factors higher than 2. Advancing the presence of a CNV on a particular exon j, based on a rule as $BF_j > 2$, should be enough to accurately indicate the true altered exons.

Once evidence on the presence of a CNV in a particular exon exists, we would want to determine its type. Using the same approach we have used to calculated the probability of a certain exonic coverage ratio lie out of the normal region, we compute the probability of a certain exonic coverage ratio lie within CNV-type regions we might be interested. Considering the genetic aspects of the CNVs, we would be interested in delimit the coverage ratio in portions that should include the $\{-1, -0.5, 1, ...\}$ CNV-types. We exclude the 1 copy type (normal homozygous), because this analysis is only opportune when evidence on the presence of a CNV exists.

	Individual 7		**Individual 9**	
exon	***BRCA1***	***BRCA2***	***BRCA1***	***BRCA2***
2	0.000	0.000	0.002	0.000
3	0.000	0.000	0.000	0.001
4	0.000	0.072	0.000	0.000
5	0.000	0.014	0.000	0.000
6	0.000	0.128	0.002	0.000
7	0.000	0.001	0.000	0.000
8	0.000	0.000	0.000	0.000
9	0.000	0.026	0.000	0.037
10	0.125	0.068	0.034	0.000
11	0.000	0.045	0.000	0.033
12	0.292	0.001	0.034	0.000
13	0.001	0.000	0.000	0.110
14	0.003	0.000	0.558	0.051
15	0.000	0.000	0.082	0.018
16	0.000	0.012	**19.370**	0.000
17	0.000	0.000	0.000	0.351
18	0.000	0.203	0.000	0.000
19	0.000	0.045	0.000	**2.403**
20	0.000	**2.789**	0.000	**5.411**
21	0.000	0.156	0.000	0.002
22	0.000	0.000	0.000	0.000
23	0.003	0.000	0.000	0.006
24		0.000		0.053
25		0.004		0.000
26		0.023		0.000
27		0.005		0.120

Table 5.2. Bayes factors calculated for the individuals 7 and 9, in both the *BRCA*1 and *BRCA*2 genes. The predictive Bayes factor was computed considering the probability of a certain exonic coverage ratio lie out of the normal region (homozygous for 1 copy): $1 - p(0.5 < y^* | y > -0.5)$. Bayes factors can be easily computed considering that y_{ij} are normally distributed.

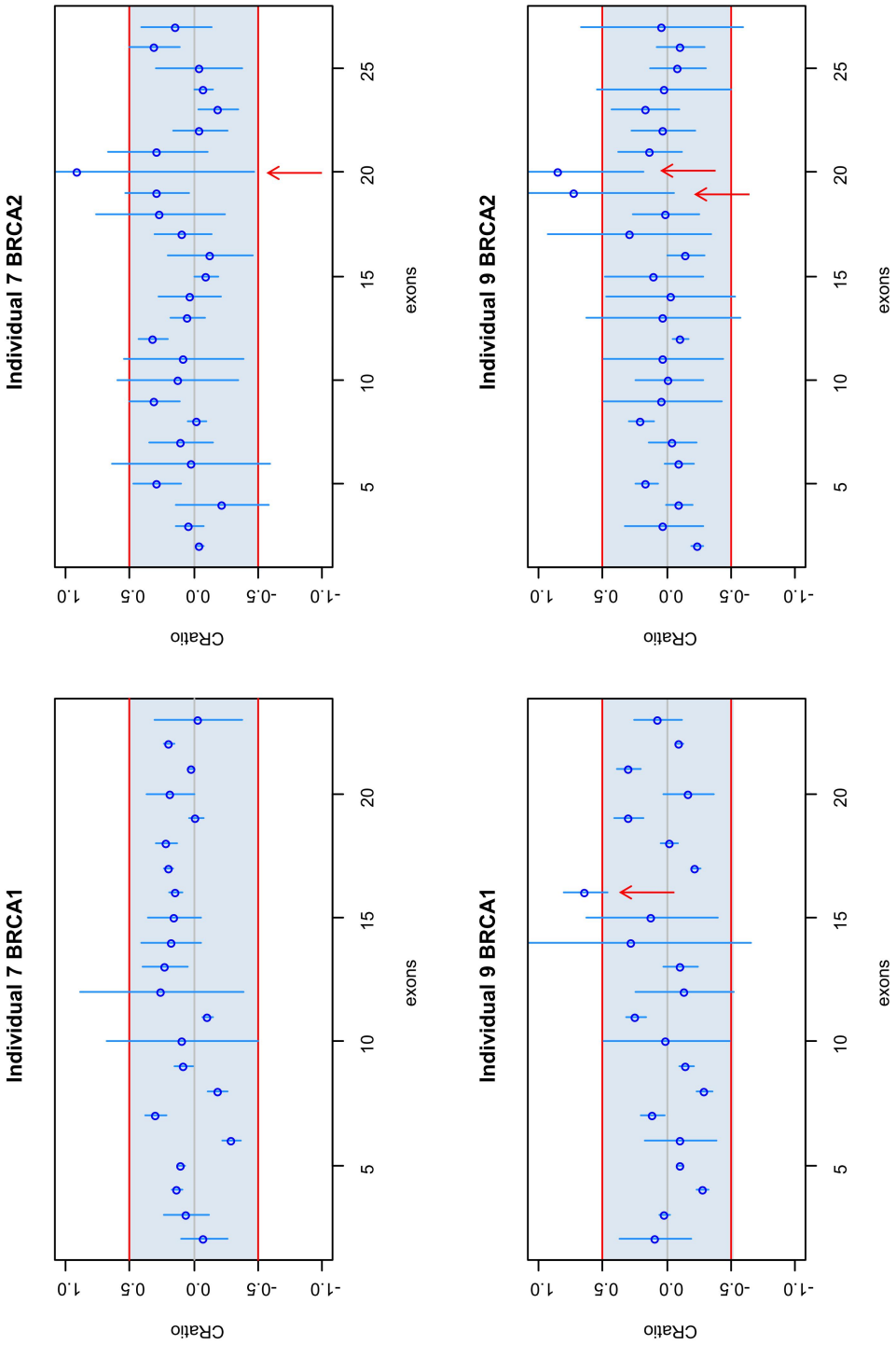


Figure 5.8. Estimates of the model parameters θ and V. θ estimates are represented by blue points and the vertical segments correspond to $2\sqrt{V}$. The blue region corresponds to the interval of the coverage ratios where the coverage readings of the case study individual follows a similar patterns as the standard individual (individual 10). Individual 7 do not possess CNVs for the *BRCA*1 gene.

The analysis of the table 5.3 shows the existence of CNVs in heterozygosity (duplication in only one of the gene copies) for all the identified exonic CNVs: the 0.5 coverage ratio region was the one that returned higher probability for the reported cases of CNV presence.

			CNV-type				
Individual	**Gene**	**exon**	**-1**	**-0.5**	**0.5**	**1**	**1.5**
7	*BRCA2*	20	0.002	0.016	**0.283**	0.256	0.136
9	*BRCA1*	16	0.000	0.000	**0.946**	0.000	0.000
9	*BRCA2*	19	0.000	0.001	**0.472**	0.213	0.023
9	*BRCA2*	20	0.000	0.000	**0.537**	0.302	0.023

Table 5.3. Probability of different CNV-types for those exons in which was obtained statistical evidence for the existence of CNVs. Non-integer values corresponds to heterozygous states, i.e. alterations that affect only one of the two gene copies in the individual's genome.

5.6.3 Final comments

The main conclusions of the *BRCA*1 and *BRCA*2 case study, considering the analysis of the model's assumptions, its predictive quality and its ability to answer to the biological problem, are:

- The data standardization using a reference coverage appears to be unaffected to the existence of coverage readings from CNV-regions. The standardization was proved to be unnecessary in this case study.

- The use of non-informative priors for the hyperparameters $(\mu, \tau, 1/\nu)$ does not affect the model's performance, and its conditional posterior distributions appear to effectively capture data information.

- The likelihood of the model shows weak signs of normality, but the posterior estimates of the model parameters θ and V are satisfactory describing the data. In both the *BRCA*1 and *BRCA*2 genes, some variances of the exonic coverage ratios were overestimated which does not seem to compromise the identification of CNVs. In a way, this feature protects the existence of false positives.

- The use of a robust distribution for the means of the exonic coverage ratios adds explanatory value to the model. In both genes we verify that the posterior estimations of ν were in the robust region of the model, suggesting that ν is a necessary and important parameter to explain data.

- An increased scale parameter (τ) appears to be a promising pattern to detect, in a preliminary phase, the presence of CNVs.
- The identification of CNVs proved straightforward using predictive exon-specific Bayes factors. Values above 2 appear sufficient to properly identify true CNVs, and most important, distinguished artificial CNVs.

Chapter 6

Conclusion

In this thesis, we advanced a robust inferential Bayesian model for the detection and identification of CNVs in protein-coding genes, considering a reference coverage. The performance of the model was evaluated in a case study involving the *BRCA*1 and *BRCA*2 genes, what gives additional insight on the statistical aspects of the model. Despite the model provided a satisfactory answer to the biological problem of CNVs identification, there are improvements which are important to discuss. Chapter 6 discusses the performance of the hierarchical robust model from an analytical and an applied point-of-view.

6.1 Reference coverage

The existence of a reference coverage prevents the use of a longitudinal standardization (expression 3.1) which is a great improvement. While it seems to be a better option, the standardization based on a reference coverage must be properly discussed (expression 3.2).

6.1.1 Limitations of the reference coverage correction

One of the limitations of the reference coverage standardization is the existence of CNV coverage readings in the tested individuals/genes. As we have already referred, during the standardization phase, we want the potential CNV coverage readings interfere as little as possible, so they can be more easily identified later using a proper statistical analysis. In the *BRCA*1 and *BRCA*2 case study, a standardization based on a linear relationship between the coverages was verified to vary little whether the coverage readings of exons with CNVs were used or excluded.

Other concern is the number of exons with CNVs. In the *BRCA*1 and *BRCA*2 case study we have a maximum of 2 exonic CNVs (*BRCA*2, individual 9), however it can be the case of a higher number. In this scenario a possible minimizing approach would be the use of the median regressor to estimate $E[C^1|C^0]$, which would be less sensitive to extreme observations. Another possible approach would be to remove a certain percentage of the most extreme observations considering a preliminary estimation of $E[C^1|C^0]$. However, this alternative could be worthless because it would be difficult to determine which percentage of points to remove and also, we observed in the case study that the CNV coverage readings are not always the outlier observations (considering the standardization line; figure 5.2).

6.2 The hierarchical Bayesian model

The hierarchical robust model was shown to be a good approach to detect and identify CNVs. There are however some distributional aspects of the model that required further discussion.

6.2.1 The hyperparameters: priors and posteriors

The hyperparameters of the model (μ, τ and ν) were assumed as having an uniform prior distribution, which is an uninformative type of prior. The marginal posterior distributions of the hyperparameters showed to be quite informative (figure 5.2), suggesting that hyperparameters are properly capturing data information. In the hierarchical model, the

hyperparameters affect the data only through the model parameters (θ and V), and the combined data of the J exons appears enough to produce informative posterior distributions.

If we had additional information on the hyperparameters or wanted to integrate information of previous analysis, we may have opted for informative priors. A prior distribution of the exponential family, would be a recommended option for μ and τ, since their conditional posteriors have both an exponential kernel (equations 4.6 and 4.7). For the robustness parameter ν it would be more difficult to define an informative prior. It must be considered that the conditional posterior distribution of this parameter is quite complex (equation 4.10), and thus, it would be challenging to find an informative prior without further analytical and numerical complications. Considering both the density and the support of the marginal posterior distribution of $1/\nu$ (figure 5.3), a Beta prior distribution appears suitable. However, it must be noted that the conditional posterior distribution of ν (equation 4.10) is not Beta-conjugated.

Note that using informative priors for the hyperparameters do not necessarily imply prior-dominated posteriors, we may (and should) chose prior parameters that result in general priors.

6.2.2 The normal likelihood

One of the limiting aspects of the hierarchical Bayesian model was the normality of the observed coverage ratios. It was found that some of the variances were overestimated. In the hierarchical model, the exon-specific coverage ratio variances are estimated as a weighted combination of the complete pool and the none at all (Gelman et al., 2014f).

$$V_j = \lambda_j s^2_{\cdot j} + (1 - \lambda_j) s^2_{\cdot\cdot}$$

It seems that the presence of some outliers in the observed coverage ratios is promoting the contribution of the $s^2_{\cdot\cdot}$ factor in the estimation of V, leading to its overestimation in some cases. That does not imply, at any rate, that we should leave the heterocedastic assumption, but instead that we should consider other likelihood models.

A natural model for the likelihood would definitely be a robust distribution, as the discussed t-student distribution. A t-student likelihood,

$$y_{ij} \sim t_\eta(\theta, \sigma^2)$$

would be able to model the variability of the observations at two levels: the variation around the mean (σ) and the existence of outlier events (η). In fact, the same approach proved useful describing the exonic coverage ratio means.

Considering the independence of the $y_{.j}$ observations and the exchangeability of the exonic coverage ratio distributions, the likelihood takes the form,

$$p(y|\theta,\sigma,\eta) = \prod_j \prod_i \frac{\Gamma\left(\frac{\eta_j+1}{2}\right)}{\Gamma\left(\frac{\eta_j}{2}\right)\sqrt{\eta_j\pi}\sigma_j}\left[1+\frac{1}{\eta_j}\left(\frac{y_{ij}-\theta_j}{\sigma_j}\right)^2\right]^{-\frac{\eta_j+1}{2}}$$

There are however some aspects to consider on the t-sample model:

- The existence of a new parameter η must be included in the model with an hierarchical dependence on the hyperparameters. A possible way is to expand the model, considering the correspondence between the variances on the normal and the t-student likelihoods.
$$V = \frac{\eta+2}{\eta}\sigma^2$$
- The likelihood does not have a simple expression. Most of the parameters which have a known conditional posterior in the normal likelihood, would have a complex expressions in the t-student likelihood, particularly the conditional posteriors of θ and σ. The prior and data kernels of θ and σ would no longer be conjugated.
- θ, τ and η would all required a Metropolis-Hastings step during the MCMC sampling. That results in $3J$ Metropolis-Hastings operations per iterate, that would certainly compromise the efficiency of the computational algorithm.

In sum, it would be difficult, if not impracticable, to develop a robust hierarchical model with a robust likelihood. The remaining option is to consider the consequences of the overestimation of some variances and adapt the inferential approach accordingly. However, we must refer they should be minimal once the accuracy of our model is good.

6.3 The classical solution revisited

In Chapter 4 a classical solution to the CNV detection and identification problem was advanced, but when considering the inferential limitations of the ANOVA, it was not explored further. It would be important, at this point, to compare the results of both approaches. We proceed to the implementation of ANOVA and Tuckey pairwise multiple comparisons at the 0.05 level.

All the analyzed genes/individuals indicate that exist at leat one exon with a differentiated θ_j (p-value less than 0.001 in all cases), thus suggesting that we should use differentiated θ_j for (one, more than one or all) the coverage ratio means. Using the classical approach to detect CNVs in a preliminary phase, would be worthless, being the level 0.05 excessively discriminant to detect different means. In the ANOVA approach the main difficulty would be to define a p-value, which would properly detect the existence of CNVs. In the Bayesian approach we take advantage of the hyperparameter estimates to identify potential statistics of interest. We found the scale of the exonic coverage ratio means a promising statistic to detect CNVs, but others could be considered, including distributional characteristics of the hyperparameters (quantiles, higher order moments, etc.).

exon 2	**0.95 lower**	**0.95 upper**	**adjusted p-value**
3	0.156	0.375	0.000
4	0.024	0.270	0.003
5	0.254	0.551	0.000
6	-0.015	0.301	0.143
7	0.071	0.316	0.000
8	0.296	0.593	0.000
9	0.151	0.396	0.000
10	0.122	0.322	0.000
11	0.166	0.361	0.000
12	0.009	0.262	0.021
13	0.130	0.402	0.000
14	0.103	0.312	0.000
15	0.225	0.453	0.000
16	-0.022	0.204	0.346
17	0.416	0.645	0.000
18	0.139	0.351	0.000
19	**0.850**	**1.082**	**0.000**
20	**0.975**	**1.210**	**0.000**
21	0.255	0.497	0.000
22	0.154	0.378	0.000
23	0.292	0.523	0.000
24	0.140	0.377	0.000
25	0.044	0.263	0.000
26	0.015	0.249	0.009
27	0.176	0.380	0.000

Table 6.1. Tuckey tests for the coverage ratio mean comparisons. Comparisons of the exon 2 with all the other exons in analysis in the individual 9, gene *BRCA*2. The p-values (Bonferroni adjusted) are based on the alternative hypothesis of the compared exonic coverage ratios are different.

We proceed with the Tuckey tests to retrieve conclusions about the θ parameter. We use the CNV coverage ratios from the individual 9 and the gene *BRCA*2, for which two CNVs have been described. Firstly, the output is a tremendous table: for a gene with J exons we would have to analyze $(J^2 - J)/2$ outputs. In our case we obtained 325 mean comparisons. Secondly, we obtain that the pairwise Tuckey tests return significant results in most of the cases. Table 6.1 shows part of the output Tuckey tests, corresponding to the 2nd exon mean comparisons. We observed that most of the exonic coverage ratio comparisons returned significant results, meaning that the coverage ratio mean of the exon 2 is different from most of the other exons (the same can be concluded analyzing the 0.95 confidence intervals). As we know, the exon 2 is not an altered exon . Again, a p-value of 0.05 is too discriminant for our purpose, and the CNV identification phase is prone to return more θ categories than actually exist.

In conclusion, the Bayesian approach contrasts with the Classical approach, mainly because the first provides a natural way for inference, in line with the meaningful hypothesis the biological problem requires. The Classical hypothesis testing introduces rigidity in the inference, that in the context of the CNV detection and identification problem is of no value.

6.4 Final remarks

- **Main objective:** To create an inferential model that permits to identify CNVs on the exonic regions of genes, using the next-generation sequencing coverage readings.

  Our main objective was accomplished since we were able to satisfactory identify and characterize exonic CNVs.

- **Objective 1.** To present a suitable strategy to deal with the coverage noise (i.e. the coverage variations that are not related with the presence of exonic CNVs).

  The coverage ratios are calculated using the next-generation sequencing technique and are thus, dependent on several experimental variables (the quantity of reagents, the DNA in the sample and the experimental design to assess the genomic regions of interest). This sources of variability are not due to the presence of CNVs, and using a coverage standardization by a reference genome, appears to effectively remove the noise introduce by the experimental variables. Indeed, we observed that coverage profiles from different individuals were proportional, feature that contributed for the easy removal of the experiment-induced variability.

- **Objective 2.** To construct a probabilistic model for the observed exonic coverage ratio modeling.

  We use the hierarchical Bayesian model for the exon coverage ratio modeling, and considering the discussion we have been doing, we were able to proper identify CNVs in protein-coding genes. Our model is overestimating some of the exonic variances, without however compromising the accuracy of the CNV identification. We should consider improvements on the normal likelihood, as the use of a robust sample model or, alternatively, the assessment of the model precision using known CNV cases, as we have done here. Notice that we found this feature not limitative for the CNV identification, since θ, the parameter that indicated the CNV-type, is well predicted by the model.

- **Objective 3.** To create an inferential framework to detect the presence of CNVs in protein-coding genes, in a early phase of the CNVs analysis.

  We have shown to be able to detect CNVs using the scale parameter of the t-student distribution. This parameter showed elevated in cases where CNVs where present. More sample will be needed however, to determine a proper threshold that distinguishes the cases we should devote a proper analysis from the normal ones.

- **Objective 4.** To develop an inferential framework to identify CNVs, when a potential altered-gene is detected. The identification of CNVs includes determining the type of exonic CNVs (deletion, duplication, etc.).

  We were able, with the use of exon-specific Bayes factors, to determine the existence of a CNVs in specific exons, and additionally, characterize the type of CNV (heterozigous, homozigous, deletions or duplications). For the *BRCA*1 and *BRCA*2 genes we were able to advance a Bayes factor threshold ($BF > 2$) that allows the identification of CNVs and, more important, the proper exclusion of artifact CNVs. This procedure may be generalized for any gene we might want to study.

Bibliography

J. Albert. *Bayesian Computation with R*, chapter Introduction to Bayesian Thinking, pages 28–34. Springer, London, 2 edition, 2009a.

J. Albert. *Bayesian Computation with R*, chapter Introduction to Bayesian Computation, page 97. Springer, London, 2 edition, 2009b.

J. Albert. *Bayesian Computation with R*, chapter Markov Chain Monte Carlo Methods, pages 117–124. Springer, London, 2 edition, 2009c.

C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. *Nature reviews. Genetics*, 12(5):363–376, 2011.

T. Bayes and R. Price. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

J. Berger and R. Wolpert. *The likelihood principle*, volume 6 of *Lecture notes – Monograph series*, chapter The likelihood principle and generalizations, pages 19–23. Institute of Mathematical Statistics, Hayward California, 2 edition, 1988.

J. Bernardo. The concept of exchangeability and its applications. *Far East Journal of Mathematical Science*, Bhattacharya Memorial Volume:1–7, 1996.

J. Bernardo. *Probability and Statistics*, chapter Bayesian Statistics, pages 1–45. Encyclopedia of Life Support Systems. A Integrated Virtual Library, 2003.

B. Carlin and T. Louis. *Bayesian Methods for Data Analysis*, chapter The Bayes approach, pages 15–98. CRC Press, United States of America, 3 edition, 2009a.

B. Carlin and T. Louis. *Bayesian Methods for Data Analysis*, chapter Bayesian computation, pages 120–158. CRC Press, United States of America, 3 edition, 2009b.

D. F. Easton. How many more breast cancer predisposition genes are there? *Breast Cancer Research*, 1(1):14, 1999.

J. D. Fackenthal and O. I. Olopade. Breast cancer risk associated with brca1 and brca2 in diverse populations. *Nature Reviews Cancer*, 7(12):937–948, 2007.

M. Fanciulli, P. J. Norsworthy, E. Petretto, R. Dong, L. Harper, L. Kamesh, J. M. Heward, S. C. L. Gough, A. de Smith, A. I. F. Blakemore, P. Froguel, C. J. Owen, S. H. S. Pearce, L. Teixeira, L. Guillevin, D. S. C. Graham, C. D. Pusey, H. T. Cook, T. J. Vyse, and T. J. Aitman. Fcgr3b copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature genetics*, 39(6):721–723, 2007.

L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nature reviews. Genetics*, 7(2):85–97, 2006.

D. Gamerman and H. Lopes. *Markov Chain Monte Carlo: stochastic simulation for Bayesian Inference*, chapter Approximate methods for Inference, pages 95–98. Chapman & Hall CRC, United States of America, 2 edition, 2006a.

D. Gamerman and H. Lopes. *Markov Chain Monte Carlo: stochastic simulation for Bayesian Inference*, chapter Gibbs sampling, pages 141–169. Chapman & Hall CRC, United States of America, 2 edition, 2006b.

D. Gamerman and H. Lopes. *Markov Chain Monte Carlo: stochastic simulation for Bayesian Inference*, chapter Metropolis-Hastings algorithm, pages 191–217. Chapman & Hall CRC, United States of America, 2 edition, 2006c.

A. Gelman. *Encyclopedia of Environmetrics*, volume 3, chapter Prior distribution, pages 1634–1637. John Wiley & Sons, Chichester, 2002a.

A. Gelman. *Encyclopedia of Environmetrics*, volume 3, chapter Prior distribution, pages 1634–1637. John Wiley & Sons, Chichester, 2002b.

A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511, 1992.

A. Gelman and C. Shalizi. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66:8–38, 2013.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*, chapter Probability and Inference, pages 3–13. CRC Press, 3 edition, 2014a.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*, chapter Basics of Markov chain simulation, pages 275–288. CRC Press, 3 edition, 2014b.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*, chapter Computationally efficient Markov chain simulation, pages 300–307. CRC Press, 3 edition, 2014c.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*, chapter Models for robust inference, pages 435–446. CRC Press, 3 edition, 2014d.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*, chapter Single-parameter models, pages 29–56. CRC Press, 3 edition, 2014e.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*, chapter Hierarchical models, pages 101–108. CRC Press, 3 edition, 2014f.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*, chapter Model Checking, pages 141–161. CRC Press, 3 edition, 2014g.

J. Ghosh, M. Delampady, and T. Samanta. *An Introduction to Bayesian Analysis Theory and Methods*, chapter Bayesian Inference and Decision Theory, pages 29–49. Springer, United States of America, 2006.

E. Gonzalez, H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez, G. Catano, R. J. Nibbs, B. I. Freedman, M. P. Quinones, M. J. Bamshad, K. K. Murthy, B. H. Rovin, W. Bradley, R. A. Clark, S. A. Anderson, R. J. O'connell, B. K. Agan, S. S. Ahuja, R. Bologna, L. Sen, M. J. Dolan, and S. K. Ahuja. The influence of ccl3l1 gene-containing segmental duplications on hiv-1/aids susceptibility. *Science*, 307(5714):1434–1440, 2005.

P. Good. Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical*, 1(2), 2002.

P. Hoff. *A First Course in Bayesian Statistical Methods*, chapter Introduction and examples, pages 3–13. Springer, London, 3 edition, 2009.

R. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

A. C. V. Krepischi, M. I. W. Achatz, E. M. M. Santos, S. S. Costa, B. C. G. Lisboa, H. Brentani, T. M. Santos, A. Goncalves, A. F. Nobrega, P. L. Pearson, A. M. Vianna-Morgante, D. M. Carraro, R. R. Brentani, and C. Rosenberg. Germline dna copy number variation in familial and early-onset breast cancer. *Breast Cancer Research*, 14(1):R24, 2012.

K. M. Kuusisto, O. Akinrinade, M. Vihinen, M. Kankuri-Tammilehto, S.-L. Laasanen, and J. Schleutker. Copy number variation analysis in familial brca1/2-negative finnish breast and ovarian cancer. *PLoS ONE*, 8(8):e71802, 2013.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25 (16):2078–2079, 2009.

D. Lindley. The philosophy of statistics. *Journal of the Royal Statistical Society Series D*, 43(3):293–337, 2000.

D. MacKay. *Information Theory, Inference, and Learning Algorithms*, chapter Efficient Monte Carlo Methods, pages 387–390. Cambridge University Press, 4 edition, 2005a.

D. MacKay. *Information Theory, Inference, and Learning Algorithms*, chapter Probabilities and Inference, pages 357–399. Cambridge University Press, 4 edition, 2005b.

M. L. Metzker. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11 (1):31–46, 2010.

M. Meyerson, S. Gabriel, and G. Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nature reviews. Genetics*, 11(10):685–696, 2010.

R. Neal. *Handbook of Markov Chain Monte Carlo*, chapter MCMC using Hamiltonian dynamics, pages 113–144. Chapman & Hall CRC Press, 2011a.

R. Neal. *Handbook of Markov Chain Monte Carlo*, chapter Inference from Simulations and Monitoring Convergence, pages 163–173. Chapman & Hall CRC Press, 2011b.

T. Pal, J. Permuth-Wey, J. A. Betts, J. P. Krischer, J. Fiorica, H. Arango, J. LaPolla, M. Hoffman, M. A. Martino, K. Wakeley, G. Wilbanks, S. Nicosia, A. Cantor, and R. Sutphen. Brca1 and brca2 mutations account for a large proportion of ovarian carcinoma cases. *Cancer*, 104(12):2807–2816, 2005.

R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, 2006.

J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y.-H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimäki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M.-C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler. Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449, 2007.

N. Sepúlveda, S. G. Campino, S. A. Assefa, C. J. Sutherland, A. Pain, and T. G. Clark. A poisson hierarchical modelling approach to detecting copy number variation in sequence coverage data. *BMC genomics*, 14:128, 2013.

P. Stankiewicz and J. R. Lupski. Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61(1):437–455, 2010.

J. Venna, S. Kaski, and J. Peltonen. Visualizations for assessing convergence and mixing of mcmc. In *Proceedings of the 14th European Conference on Machine Learning (ECML 2003)*, pages 432–443, Berlin, 2003. Springer.

M. Zhao, Q. Q. Wang, P. Jia, and Z. Zhao. Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics*, 14(11):1–15, 2013.

Appendix A

Computational implementation

During the development of this thesis we developed R codes with the objective of implement the hierarchical Bayesian model. The outline of the code organization is represented in figure A.1. The code was developed in four phases.

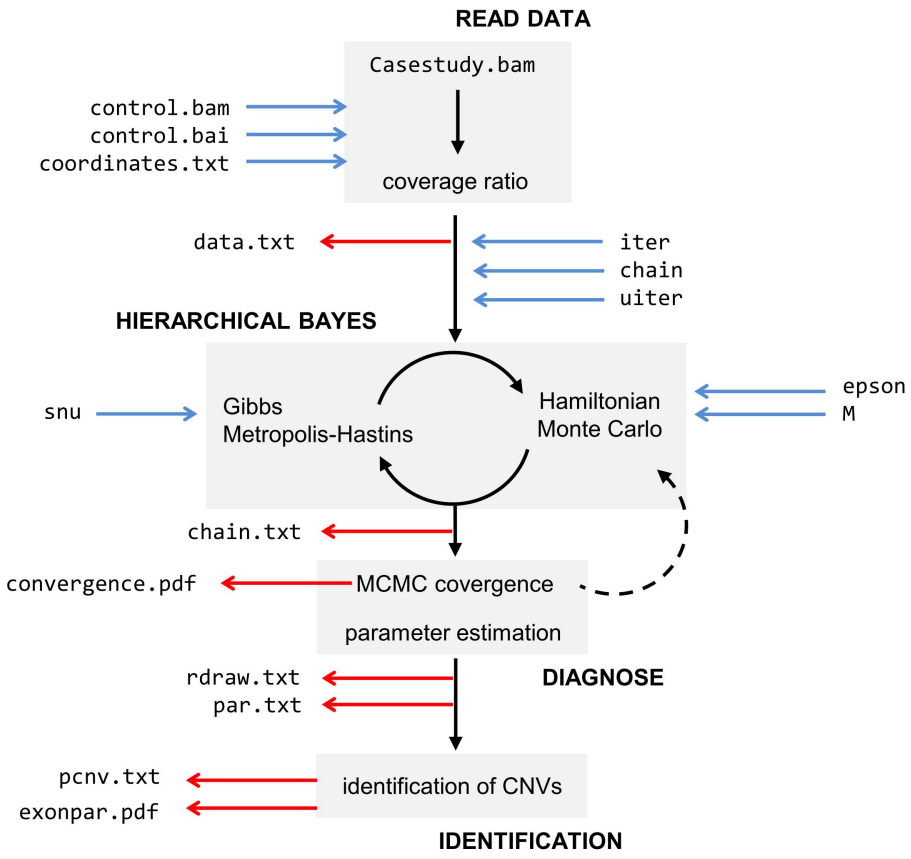


Figure A.1. Computation implementation outline. The input and output parameters and/or files are represented in blue and red arrows, respectively.

The implementation of the hierarchical Bayesian model requires the definition of several parameters, namely, the number of iterates (`iter`), the number of chains (`chains`) and the number of utile iterations (`uiter`). The `iter/uiter` ratio is, by definition, the thinning fraction. Other parameters are related with the MCMC part of the algorithm, as the standard deviation of the proposal distribution (`snu`), the leapfrog discretization factor (`epson`) and the mass matrix (`M`).

```
iter   <- 30000
chains <- 2
uiter  <- 1500
snu    <- 0.1
epson  <- 0.1
M      <- rep(1,D)
```

The algorithm begins by extracting the data from the `.bam` and `.bai` files from the case study and the reference individuals. A coordinate file, with the information of the genes to analyze and the genomic coordinates to extract the coverage, must exist in the working directory. The algorithm proceeds with the calculation of the coverage ratio for each of the case study individual in the working directory.

```
#READ DATA

#PACKAGES
library(rbamtools)

path  <- getwd()

#COORDINATE FILE
coord <- read.table(paste(path,"coordinates.txt",sep="/"),header=T)
gene  <- as.vector(unique(coord[,1]))

genei <- matrix(ncol=4,nrow=length(gene))
for (i in 1:length(gene)) {
  genei[i,1:4] <- c(gene[i],coord[which(coord[,1]== gene[i])[1],2],
                    coord[which(coord[,1]== gene[i])[1],3],
                    coord[which(coord[,1]== gene[i])[1],4])
}

#CONTROL.BAM
control <- bamReader(paste(path,"control.bam",sep="/"))
load.index(control,paste(paste(path,"control.bam",sep="/"),"bai",sep="."))
index.initialized(control)

files <- list.files(path)
files <- files[which(substr(files,nchar(files)-2,nchar(files))=="bam")]
```

```
files <- files[-which(files == "control.bam")]  #all the files in the directory
file  <- substr(files,1,nchar(files)-4)

for (j in 1:length(files)) {
  sample <- bamReader(paste(path,files[j],sep="/"))
  load.index(sample,paste(paste(path,files[j],sep="/"),"bai",sep="."))
  index.initialized(sample)
  rdf <- getRefData(sample)

  fpath <- paste(path,file[j],sep="/")
  dir.create(fpath)

  for (i in 1:length(gene)) {
    gpath <- paste(fpath,gene[i],sep="/")
    dir.create(gpath)

    #EXONS
    exoni <- coord[which(coord[,1] == gene[i]),5:7]

    #COVERAGE
    scov <- c()
    ccov <- c()
    exon <- c()

    for (k in 1:dim(exoni)[1]){
      exon  <- c(exon,rep(exoni[k,1],exoni[k,3]-exoni[k,2]+1))
      exonc <- as.integer(c(as.numeric(genei[i,2])-1,exoni[k,2],exoni[k,3]))
      scov  <- c(scov,getDepth(alignDepth(bamRange(sample ,exonc))))
      ccov  <- c(ccov,getDepth(alignDepth(bamRange(control,exonc))))
    }

    ycov <- scov/1000
    xcov <- ccov/1000
    a    <- sum(ycov*xcov)/sum(xcov*xcov)

    rcov <- scov/(a*ccov) - 1

    write.table(cbind(exon,rcov),paste(gpath,"edata.txt",sep="/"),row.names=F,quote=F)

    data <- matrix(NA,ncol=4,nrow=dim(exoni)[1])
    for (k in 1:1:dim(exoni)[1]){
      ercov <- rcov[which(exon == exoni[k,1])]
      data[k,1:4] <- c(exoni[k,1],mean(ercov),var(ercov),length(ercov))
    }

    colnames(data) <- c("exon","mj","s2j","nj")
    write.table(data,paste(gpath,paste("data.txt"),sep="/"),row.names=F,quote=F)
  }
}
```

Once the data is extracted, the algorithm implements the MCMC step. It includes the Gibbs-Metropolis-Hastings and the Hamiltonian Monte Carlo algorithms that proceed iteratively and produce a `chain.txt` file, which contains `usim` iterates for each of parameters in analysis.

```
#HIERARCHICAL BAYES

#FUNCTIONS
rnu <- function(nu,tau,V,snu) {
  inu <- rnorm(1,1/nu,snu)
  if (inu <= 0 | inu >1) {
    jp <- 0
  } else {
    nus <- 1/inu
    r   <- exp( nus*( J*log(nus/2)/2 + J*log(tau) -
                sum(log(V))/2 - tau*tau*sum(1/V)/2 ) - J*lgamma(nus/2) -
                nu*( J*log(nu /2)/2 + J*log(tau) -
                sum(log(V))/2 - tau*tau*sum(1/V)/2 ) + J*lgamma(nu /2)  )
    if (runif(1) < r) {
      nu <- nus
    }
    jp <- min(r,1)
  }
  return(c(nu,jp))
}

lpost <- function(lambda,mj,s2j,nj) {
  J     <- length(lambda)/2-1.5
  mu    <- lambda[1]
  tau   <- exp(lambda[2])
  theta <- lambda[3:(2+J)]
  V     <- exp(lambda[(3+J):(2*J+2)])
  nu    <- exp(lambda[2*J+3])
  lpost <- J*nu*log(nu/2)/2 - J*lgamma(nu/2) + nu*J*log(tau) -
           sum((nu*tau^2 + (theta-mu)^2 + nj*(mj-theta)^2 + (nj-1)*s2j )/V)/2 -
           sum((nu+3+nj)*log(V))/2 +
           log(tau) + sum(log(V)) + log(nu)
  return(lpost)
}

lgrad <- function(lambda,mj,s2j,nj) {
  J     <- length(lambda)/2-1.5
  mu    <- lambda[1]
  tau   <- exp(lambda[2])
  theta <- lambda[3:(2+J)]
  V     <- exp(lambda[(3+J):(2*J+2)])
  nu    <- exp(lambda[2*J+3])
  grad  <- c( sum((theta-mu)/V),
              nu*J - nu*tau*tau*sum(1/V) + 1,
```

```
              (mu + nj*mj - (1+nj)*theta)/V,
              (nu*tau*tau+(theta-mu)^2+nj*(mj-theta)^2+(nj-1)*s2j)/(2*V) - (nu+3+nj)/2 + 1,
              nu*(J*log(nu/2)/2 + J/2 - J*digamma(nu/2)/2 + J*log(tau) - sum(log(V))/2 -
              tau*tau*sum(1/V)/2 ) + 1 )
  return(grad)
}

hmc <- function(lambda,M,epson,mj,s2j,nj){
  L  <- floor(1/epson)-1
  D  <- length(lambda)

  #RANDOM MOMENTUM
  psi     <- rnorm(D,0,sqrt(M))
  lambdai <- lambda
  lposti  <- lpost(lambda,mj,s2j,nj) - sum(psi*psi/M)/2

  #LEAPFROG STEP
  psi    <- psi + epson*lgrad(lambda,mj,s2j,nj)/2
  for (i in 2:(L-1)) {
    lambda <- lambda + epson*psi/M
    psi    <- psi + epson*lgrad(lambda,mj,s2j,nj)
  }
  psi <- psi + epson*lgrad(lambda,mj,s2j,nj)/2

  #METROPOLIS STEP
  r <- exp(lpost(lambda,mj,s2j,nj) - sum(psi*psi/M)/2 - lposti)
  if (is.na(r) == T) { r <- 0 }
  if ( runif(1) <= min(1,r) ) {
    return(c(1,lambda))
  } else {
    return(0)
  }
}

#HIERARCHICAL ROBUST MODEL
for (j in 1:length(files)) {
  for (i in 1:length(gene)) {

    gpath <- paste(path,file[j],gene[i],sep="/")
    data  <- read.table(paste(gpath,"data.txt",sep="/"),header=T)

    #STATISTICS
    mj  <- data[,2]
    s2j <- data[,3]
    nj  <- data[,4]
    J   <- dim(data)[1]

    D    <- 2*J+3
```

```
    #RUN
    usims    <- seq(iter/uiter,iter,iter/uiter)

    for (k in 1:chains){

      nu  <- 1/runif(1,0,1)
      mu  <- rnorm(1,mean(mj),sd(mj))
      tau <- runif(1,0,2*sd(mj))
      V   <- runif(J,0,2*mean(s2j))

      p     <- 1
      gibbs <- matrix(NA,ncol=4+J*2,nrow=uiter)

      for (m in 1:iter){

        theta <- rnorm(J,(mu+mj*nj)/(nj+1),sqrt(V/(nj+1)))
        V     <- (nu*tau^2+(theta-mu)^2+nj*(mj-theta)^2+(nj-1)*s2j)/rchisq(J,nu+nj+1)
        mu    <- rnorm(1,sum(theta/V)/sum(1/V),sqrt(1/sum(1/V)))
        tau   <- sqrt(rgamma(1,J*nu/2+1,nu*sum(1/V)/2))
        sam   <- rnu(nu,tau,V,snu)
        nu    <- sam[1]

        phmc <- hmc(lambda=c(mu,log(tau),theta,log(V),log(nu)),M,epson,mj,s2j,nj)
        if (phmc[1] == 1) {
          lambda <- phmc[-1]
          mu     <- lambda[1]
          tau    <- exp(lambda[2])
          theta  <- lambda[3:(2+J)]
          V      <- exp(lambda[(3+J):(2*J+2)])
          nu     <- exp(lambda[2*J+3])
        } else {
        }

        if (m == usims[p]) {
          gibbs[p,] <- c(mu,tau,theta,V,1/nu,sam[2])
          p <- p+1
        }
      }
      colnames(gibbs) <- c( "mu","tau",paste("theta",data[,1],sep=""),
                                       paste("V"    ,data[,1],sep=""),"inu","jp" )
      write.table(gibbs,paste(gpath,paste("chain",k,".txt",sep=""),sep="/"),row.names=F,quote=F)
    }
  }
}
```

During the diagnose step the MCMC iterates are visually and numerically analysed. For each chain, the first 1/3 of the iterates are eliminated. The remaining iterates are evaluated for convergence, mixing and correlation. Once these criteria are respected a file with the converged, mixed and independent iterates is produced (`rdraw.txt`) along with the parameters estimates according to the quadratic loss criteria (mean and standard deviation).

```
#DIAGNOSE
for (j in 1:length(file)) {
  for (i in 1:length(gene)) {
    gpath <- paste(path,file[j],gene[i],sep="/")

    cfile  <- list.files(gpath)
    cfile  <- cfile[which(substr(cfile,1,5)=="chain")]
    chains <- length(cfile)

    #CONVERGENCE AND MIXING
    rdraw <- NULL
    for (k in 1:chains) {
      cdraw <- as.matrix( read.table(paste(gpath,paste("chain",k,".txt",sep=""),sep="/"),
                                                    header=T) )
      rdraw  <- rbind(rdraw,cdraw[501:1500,])
    }
    write.table(rdraw,paste(gpath,"rdraw.txt",sep="/"),row.names=F,quote=F)

    it <- 1:1000
    pdf(paste(gpath,"convergence.pdf",sep="/"),width = 6, height = 4.5)
    par(mfrow=c(1,3))
    for (k in c(1,2,dim(rdraw)[2]-1)) {
      plot(it,rdraw[it,k])
      for (l in 2:chains) {
        points(it,rdraw[it+(l-1)*1000,k],col=l)
      }
    }
    dev.off()

    #PARAMETERS
    par <- t(rbind(apply(rdraw,2,mean),apply(rdraw,2,sd)))
    colnames(par) <- c("mean","sd")
    write.table(par,paste(gpath,"par.txt",sep="/"),row.names=T,quote=F)

  }
}
```

The identification phase includes the visual inspection of the exonic parameters (θ and V) and the calculation of the predicted Bayes factors for each exon. Two output files are produced, respectively: `exonpar.pdf` and `pcnv.txt`.

```
#INDENTIFICATION
for (j in 1:length(file)) {
  for (i in 1:length(gene)) {

    li <- c(-0.5,0.5)
    gpath <- paste(path,file[j],gene[i],sep="/")

    data   <- read.table(paste(gpath,"data.txt",sep="/"),header=T)
    param  <- read.table(paste(gpath,"par.txt",sep="/"),header=T)
    J      <- dim(param)[1]/2 - 2
    theta  <- param[3:(J+2),1]
    V      <- param[(J+3):(2*J+2),1]

    int <- data[,1]
    v   <- sqrt(V)

    pdf(paste(gpath,"exonpar.pdf",sep="/"),width = 6.5, height = 4)
    plot(int,rep(-100,J),ylim=c(li[1]-0.5,li[2]+0.5),xlab="exons",ylab="CRatio",
              main=paste("Individual ",j," ",gene[i],sep=""))
    abline(h=0,col="gray")
    abline(h=c(li[1],li[2]),col="red")

    segments(x0=int,y0=theta-2*v,x1=int,y1=theta+2*v,col="dodgerblue")
    points(int,theta,col="blue")
    dev.off()

    #PRED BAYES FACTOR
    rdraw  <- as.matrix( read.table(paste(gpath,"rdraw.txt",sep="/"),header=T))

    rtheta  <- rdraw[1:2000,3:(J+2)]
    rv      <- sqrt(rdraw[1:2000,(J+3):(2*J+2)])

    pcnv <- matrix(NA,ncol=2,nrow=J)
    for (k in 1:J){
      pred     <- rnorm(2000,rtheta[,k],rv[,k])
      prob     <- 1 - pnorm(li[2],mean(pred),sd(pred)) + pnorm(li[1],mean(pred),sd(pred))
      pcnv[k,] <- c(int[k], prob/(1-prob))
    }
    colnames(pcnv) <- c("exon","pbayesfactor")
    write.table(pcnv,paste(gpath,"pcnv.txt",sep="/"),row.names=F,quote=F)
  }
}
```